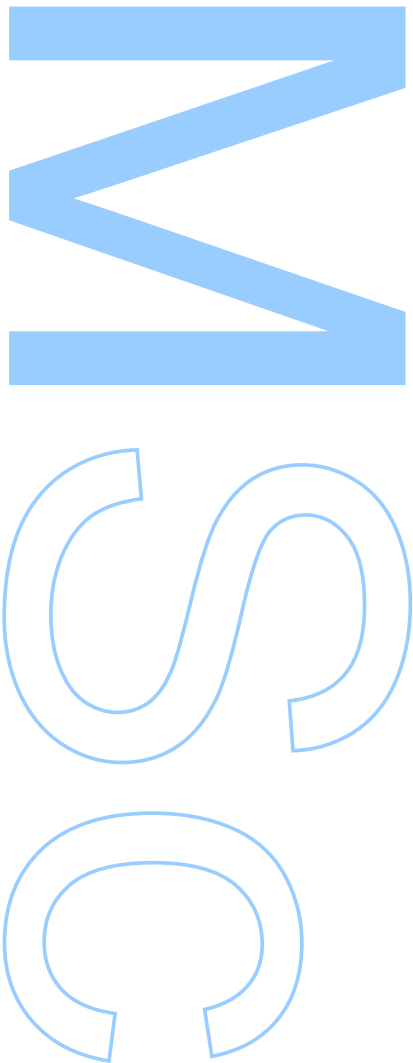
Automatic Detection of Anatomical Structures and Breast Cancer Diagnosis on X-Ray Mammography

Hugo Manuel Soares Oliveira
Master's degree in Computer Science
Computer Science Department
2018

Supervisor
Hélder Filipe Pinto de Oliveira, Phd, Faculty of Sciences, University of Porto

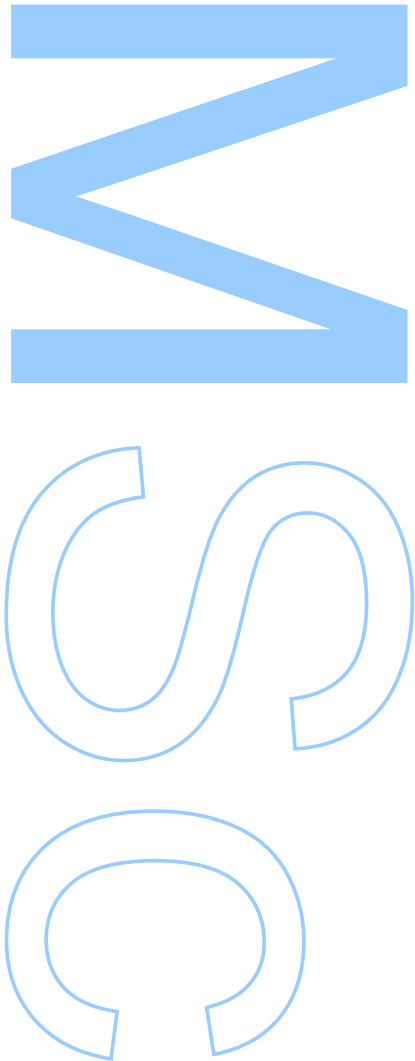
Co-Supervisor
João Pedro Fonseca Teixeira, Msc, INESC TEC



All corrections determined by the jury, and only those, were incorporated.

The President of the Jury,

Porto, _____/_____/_____



Acknowledgements

I would like to thanks to Hélder Oliveira and João Teixeira for their support and motivation. Their collaboration was helpful and we always had interesting and fruitful discussions on the emerging ideas on this subject.

I would like to thank INESC TEC for its support in using their resources, namely the GPU servers that enabled to speed up results and evaluate new model architectures.

Abstract

Breast cancer is one of the most common manifestations of cancer among women. It is estimated that 520 thousand deaths are caused by this disease each year. In order to reduce the number of clinical professionals required and the rate of false negatives and false positives, the scientific community made efforts to develop effective computed aided diagnoses Computer Aided Systems (CAD) systems to assist specialist while increasing the accuracy of the final diagnosis. While the number of false positives attains in the increase of unnecessary clinical trials, the false negatives must be zero. For this task, several authors proposed the use of computer vision and deep learning techniques to assist diagnosis.

In this work we study and implement the main stages of a complete pipeline in the context of breast cancer CAD system to identify suspicious regions while discarding nonrelevant ones, characterize the positively identified lesions and assess its severity using machine learning methods. Each of the stages is studied in detail, providing comparisons between different approaches.

First, mammograms images are subject to a pre-processing to remove unwanted regions that don't provide useful information about lesion manifestations and can degrade the system performance.

Second, suspicious lesions are characterized to obtain information about the lesion characteristics.

Third, suspicious lesion regions are characterized in order to asses its severity. Traditional computer vision and deep learning approaches were employed in this task. For traditional computer vision methods, several models and strategies for binary and multi-class classification were evaluated. Concerning deep learning models, a large dataset was constructed from breast images, enabling deep learning models to focus on the differences between lesions/background regions while increasing the amount of data available for training. A cascade configuration using Convolutional Neural Networks (CNN) was set to detect and classify lesions.

Keywords— Breast Cancer, CAD, Screening, Lesions, Classification

Resumo

Cancro da mama é uma das mais comuns manifestações de cancro no género femenino. É estimado que 520 mil mortes sejam causadas por esta doença em cada ano. Para reduzir o numero de profissionais de saúde envolvidos e a taxa de falso negativo. A comunidade científica tem desenvolvido mecanismos automáticos de diagnóstico de forma a ajudar os profissionais e aumentar a precisão do diagnóstico final. Para a tarefa, diferentes autores propuseram o uso de técnicas de processamento de imagem combinadas com mecanismos de aprendizagem automática para diagnostico automático.

Neste trabalho, estudou-se e implementou-se as principais componentes de uma cadeia completa no contexto de um sistema CAD para identificar regiões suspeitas ao mesmo tempo que as regiões não relevantes sõo descartadas, caracterizar as lesões e avaliar a sua severidade atravês do uso de métodos de aprendizagem automtica. Cada um dos elementos dos sistema estudados em detalhe, permitiram obter comparações entre diferentes abordagens.

Inicialmente as imagens de mamogramas são sujeitas a pré-processamento para remover regiões não relevantes que não providenciam informação útil e podem degradar a performance do sistema.

Segundo, regiões suspeitas são detectadas e caracterizadas de forma obter informação sobre as características da lesão e avaliar a sua severidade.

Terceiro, regiões suspeitas e imagens são classificadas através de métodos tradicionais e redes profundas para aprendizagem automática. No que respeita aos métodos tradicionais, diferentes modelos e estratégias para classificação binária e multi-classe são avaliadas. No que se refere a redes neuronais, um vasto dataset foi construído das imagens iniciais, permitindo que os modelos se foquem nas regiões das lesões ao mesmo tempo que se aumenta o número de dados disponíveis para treino do modelo. Foi desenvolvida uma configuração em cascata usando redes convolucionais para realizar detecção, segmentação e classificação e comparada com métodos de aprendizagem automática tradicionais.

Keywords— Cancro Mama, CAD, Rastreamento, Lesões, Classificação

Acronyms

ACM Active Contours Models. 66

ANN Artificial Neural Networks. 129, 130, 135

AOM Area Overlap Measure. 47, 48, 104, 159, 175

AUC Area Under ROC curve. 49, 52, 55, 56, 93

BCDR Breast Cancer Data Repository. 52

BI-RADS Breast Imaging Reporting And Data System. 20, 26, 35, 50, 51, 58, 59, 111, 113, 114, 119, 121, 124–129, 140, 150, 151, 153, 157, 159, 160

CAD Computer Aided Systems. 5, 35–37, 44, 50, 57, 61, 105, 155

CC Craniocaudal Mammogram. 34, 46, 55

CLAHE Contrast-limited adaptive histogram equalization. 40, 41, 62

CM Combined Measure. 104, 159, 175

CNN Convolutional Neural Networks. 5, 35, 43, 55, 58, 71, 72, 86, 93, 107, 130–133, 139–141, 144, 145, 148, 151, 161

CV Cross Validation. 122, 126

DAG Directed Acyclic Graph. 48, 95

DBn Deep Belief Network. 55

DC Dice Coefficient. 43, 76, 104, 159, 176

DDSM Digital Database for Screening Mammography. 45–47, 51, 52, 55

DP Dynamic Programming. 47, 71

FN False Negatives. 31, 35, 36, 177

FP False Positives. 26, 31, 35, 36, 41, 46, 48, 49, 58, 78, 84, 86–88, 94, 144, 146–149, 153, 156, 160, 161, 177

FPR False Positive Rate. 178

GPU Graphical Processing Unit. 74

GT Ground Truth. 23, 47, 59, 60, 74, 77, 84, 86, 155–157, 160, 162, 175

HD Hausdorff Distance. 43, 159, 176

HGD Histograms of Gradient Divergence. 52

kNN k-Nearest Neighbours. 49, 51, 52, 126

LDA Linear Discriminant Analysis. 49

MAE Mean Absolute Error. 125, 126, 128, 152, 153, 157, 160, 161, 179

MIAS Mammographic Image Analysis Society. 46, 49, 51

MLO Mediolateral Oblique Mammogram. 34, 46, 55

MRI Magnetic Resonance Imaging. 30

MSE Mean Square Error. 179

NB Naive Bayes. 116, 121, 122, 124

PCA Principal Component Analysis. 108, 109, 111, 160

RANSAC Random Sample Consensus. 41

ReLU Rectified Linear Unit. 19, 20, 73, 75, 132, 133, 142

RF Random Forest. 55, 117, 121–124, 126, 127, 135, 160

ROC Receiver Operating Characteristic. 178

ROI Region of Interest. 34, 41, 47, 52, 77, 84, 95, 97, 159, 160

SGD Stochastic Gradient Descent. 73

SMOTE Synthetic Minority Over-sampling Technique. 126, 127

SP Shortest Path. 42, 48, 58, 70, 71, 75, 76, 78, 94, 95, 103, 105, 147–150, 155, 156, 159, 160

STD Standard Deviation. 136

SVM Support Vector Machine. 46, 49, 51, 52, 78, 84, 86, 87, 115, 116, 121, 122, 125, 126

Tanh Hyperbolic Tangent. 20, 142

TN True Negatives. 177

TP True Positives. 49, 88, 147, 148, 160, 161, 177

TPR True Positive Rate. 178

US UltraSound. 30

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Breast Cancer in the Society

Breast cancer is considered a massive health problem worldwide. Statistics have shown that it is accountable for 15% of cancer deaths among females between 40 and 55 years of age (Lan et al., 2012). Despite the fact that the number of breast cancer incidents has increased over the years, its early detection combined with adequate therapeutics increased the survival rate (Torre et al., 2015) by large margin. Notwithstanding, breast cancer symptoms do not appear in early ages (Nithya and Santhi, 2011; Tang, 1998). The world cancer incidence rate by country is represented in Figure 1.1.

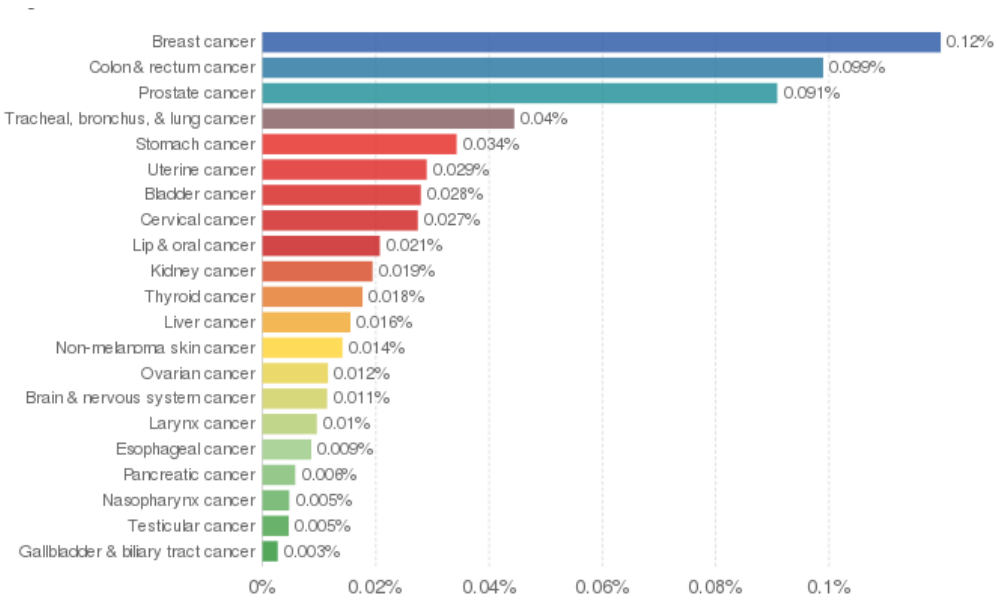


Figure 1.1: Share of population suffering from cancer types (Image from [1]).

The study in 1 shows that cancer incidence rate is unevenly distributed, presenting higher prevalence in developed regions. Although some risk factors are identified, such as age and family breast cancer history (Nithya and Santhi, 2011), is still not clear why every year more than 1 million breast cancer cases are discovered and over 400 thousand women succumb to the disease (Hela et al., 2013). In addition, many countries in South America, Africa, and Asia have witnessed the increase of the breast cancer episodes, providing strong indications that this problem will become more frequent in near future, due to the advent of the cultural and economic transition occurring in emergent countries.

Adding to the social impact, there is also a big economic burden associated with the disease, namely the required infrastructures and clinical professionals combined with the loss of productivity in the form of morbidity and mortality. According to to Blumen et al. (2016), the cost per patient in the United States is 80 715 $ for the first year after diagnosis, followed by an additional 20 822 $ for the next years.

Early diagnosis is the most effective form to reduce the mortality by breast cancer enabling to improve the survival rates to a great extent (Hela et al., 2013). For breast cancer detection, the widely accepted method for screening breast cancer corresponds to the use of low dosage X-ray (Tang, 1998; Hela et al., 2013) to obtain mammogram images. After suspicious lesion findings, further studies can be carried out by employing UltraSound (US) or Magnetic Resonance Imaging (MRI) methods to validate the findings by gathering more detailed information.

[1] https://ourworldindata.org/cancer

Screening has the purpose to collect the majority of information, enabling subsequent studies to be directed in order to provide more specificity about the diagnosis. The screening mammography is performed in asymptomatic population over regular periods to identify early signs of breast cancer such as masses, calcification's, bilateral asymmetry or architectural distortion. Further diagnoses are performed on patients that have presented abnormal clinical findings. For the screening task, different systems can be employed, however, the vast majority of mammograms are obtained by systems similar to the presented Figure 1.2.

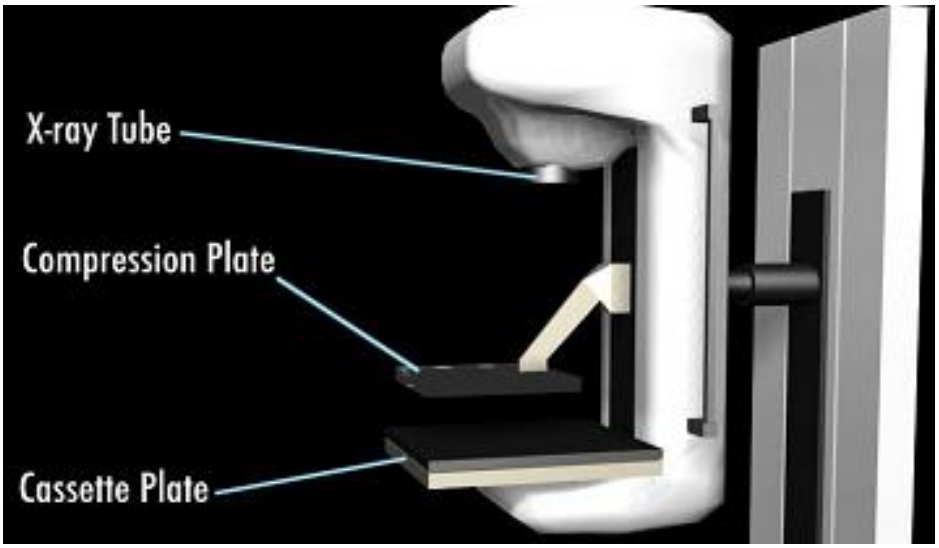


Figure 1.2: Breast cancer image system (Image from [2]).

Both screening and diagnostic are traditionally performed by radiologist by the visual inspection of the mammograms. A typical mammogram image exhibits normal structures such as fat, fibroglandular tissue, breast ducts, and nipples, including possible abnormalities. Mammograms with exception of the fat, (glands, connective tissue, and abnormalities) exhibit a high and uniform illumination, making difficult the distinction between normal and abnormal tissue (Ganesan et al., 2013a; Hela et al., 2013). As consequence, the manual screening provides vital cues that can be missed during scan study. In fact, studies have shown that mammograms are susceptible to high percentages of False Positives (FP) and False Negatives (FN). This becomes particularly problematic when radiologist classifies malignant cases as benign. The FN cases can lead to a shift in the best treatment interval, potentially endangering the patient. On the other hand, FP cases result in the allocation of unnecessary resources and treatment procedures. Statistical studies shown that radiologists classify between 10 % to 30 % malignant cases as benign (Sampat et al., 2005b; Ganesan et al., 2013a). To overcome this problematic, double reading by different radiology specialist have been advocated by the majority of the countries. This reduced the number of FN cases, however, the double reading increase the workload of scarce human resources in the

[2]https://www.cancerquest.org/patients/detection-and-diagnosis/mammography

radiologist field and the outcomes are still susceptible to human error.

1.2 Breast Cancer Physiology and X-ray Imaging

The female breast organ is particularly interesting since it serves as main nutrition support for early infants. Its structure suffers substantial changes in adulthood and has a particular cultural, social and personal relevance. Those interest combined with the high death risk associated with breast cancer episodes influences the importance of correct diagnosis and subsequent treatment (Drake et al., 2009).

Cancer is an umbrella term for a group of diseases caused by abnormal cell growth in different parts of the body. The accumulation of extra cells usually forms a mass of tissue called a tumor. Tumors can be grouped into benign or malignant: benign tumors are noncancerous, lacking the ability to invade surrounding tissue and will not regrow if removed from the body; malignant or cancerous tumors are harmful, can invade nearby organs and tissues (invasive cancer), can spread to other parts of the body (metastasis) and will sometimes regrow even when removed. The two most common types of breast cancer are ductal carcinoma and lobular carcinoma, starting in the breast ducts and lobules, respectively (Figure 1.3).

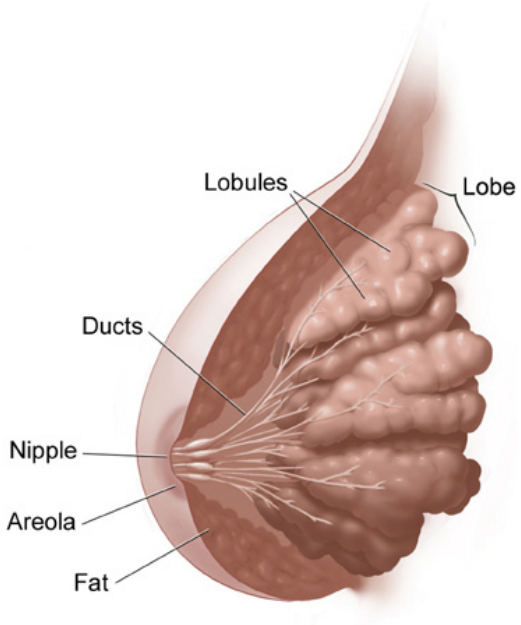


Figure 1.3: Anatomy of the female breast. (Image from [3])

The cancer stage depends on the size of the tumor and whether the cancer cells have spread to neighboring tissue or other parts of the body. Stages are expressed as a Roman numeral

[3] http://www.cancer.gov/publications/patient-education/WYNTK_breast.pdf

ranging from 0 through IV with stage I being considered early-stage breast cancer and stage IV cancer an advanced one. Stage 0 describes non-invasive breast cancers, also known as carcinoma in situ. Stage I, II and III describe invasive breast cancer, i.e., cancer has invaded normal, surrounding breast tissue. Stage IV is used to describe metastatic cancer, i.e., it has spread beyond nearby tissue to other organs of the body.

To obtain the mammogram, the breast is exposed to a short X-ray pulse which travels through the breast, being captured by a detector. The final image is obtained based on the energy absorbed by each of the breast sections. Adipose tissue is seen as dark (radio transparent) while epithelial tissue such as breast muscle and lesions are highlighted (Figure 1.4).

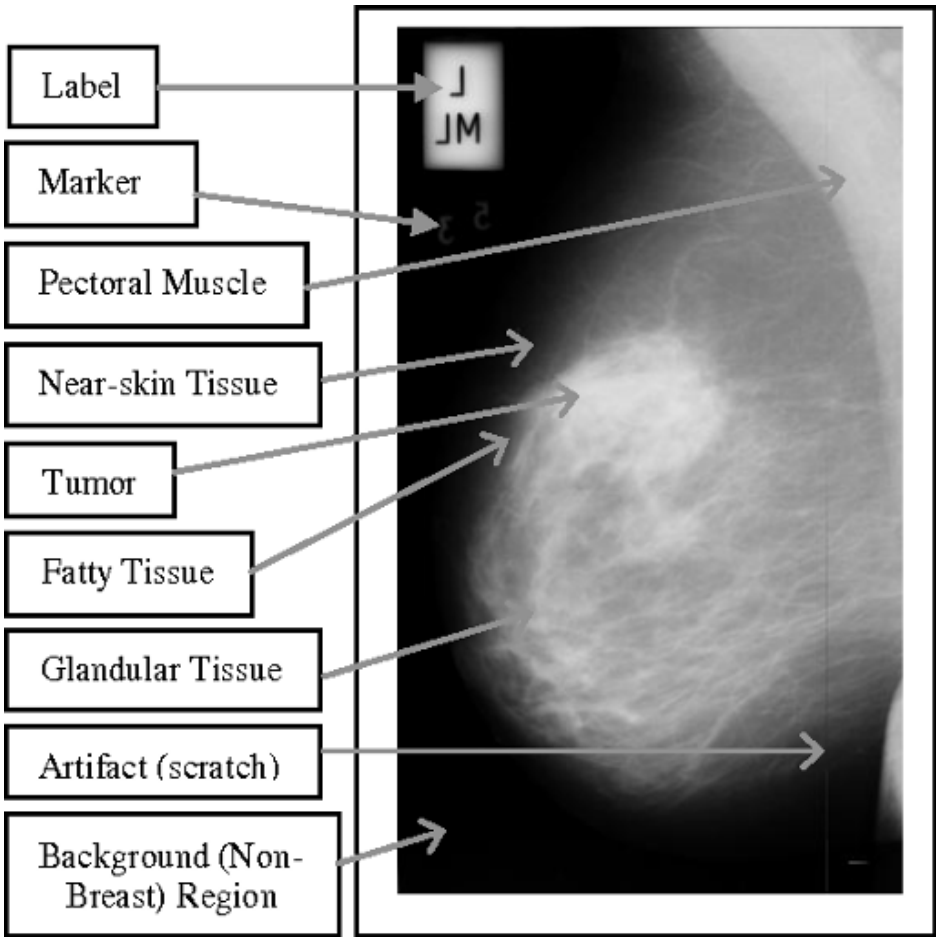
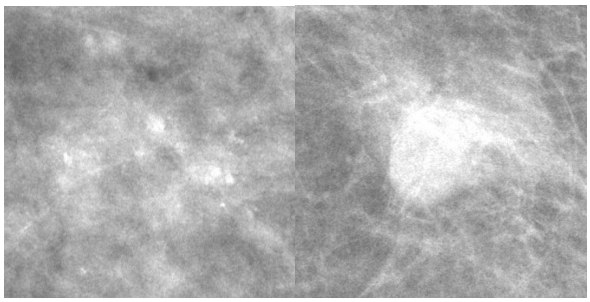


Figure 1.4: Example of an labeled breast mammogram (Image from Saidin et al. (2012)).

Radiologists look primarily for microcalcifications and breast masses. Microcalcifications are characterized by tiny deposits of calcium in the breast tissue that can be a sign of early breast cancer if found in clusters with irregular layout and shapes. Breast masses or breast lumps are a variety of things: fluid-filled cysts, tissues, noncancerous or cancerous tumors, among others. A mass can be a sign of breast cancer if it has an irregular shape and poorly

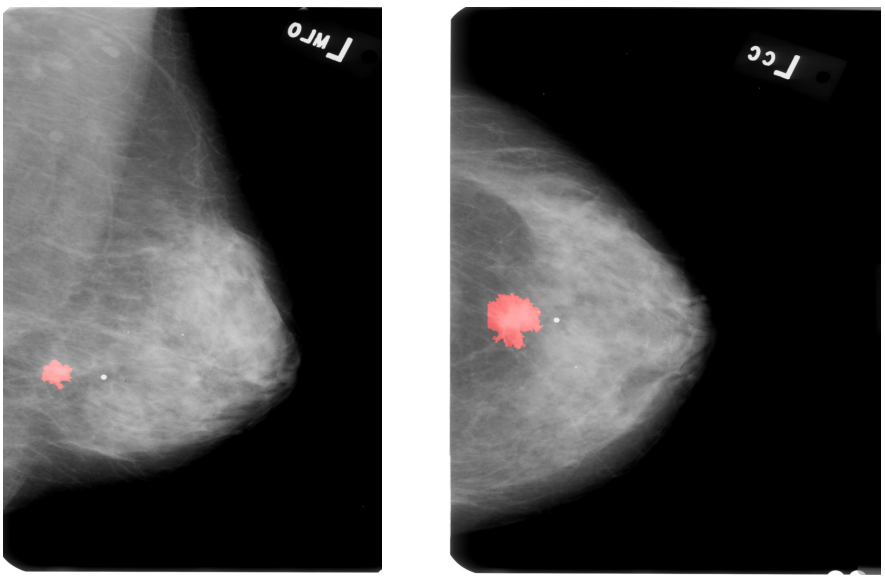
defined margins (Giger, 2014). Figure 1.5 presents an example of both cases.



Cluster Calcification Mass

Figure 1.5: Examples of cluster of calcification's and masses.

Architectural distortions can be also present and belong to one of the benign or malignant Hela et al. (2013) categories. Vast majority mammogram images are taken from two different views of the breast, Craniocaudal Mammogram (CC) and Mediolateral Oblique Mammogram (MLO), (Figure 1.6) with vertical and lateral orientations, respectively. The identified regions are defined as Region of Interest (ROI).



MLO view CC view

Figure 1.6: Two views of the same exam. Red region highlight the same mass lesion. Image from CBIS-DDSM (Lee et al., 2016)

Conventional mammography records mammograms in film, while digital mammography converts x-rays into electrical signals and stores to be stored electronically. Digital mammograms offer a clearer picture of the breast and facilitate manipulation and sharing between health care professionals.

In recent years, the community accepted as common practice an extension of the concept of benign/malignant classification. The extension in practice ranks the mammogram images, by measuring the severity of the findings while providing recommendations for each of the levels. The ranking is denominated Breast Imaging Reporting And Data System (BI-RADS) and ranges from $[0, 6]$, summarized in Table 1.1.

Table 1.1: Bi-RADS Categories, interpretation and recommended actions.

Category	Interpretation	Recommendations
0	Insufficient study	Obtain additional imaging
1	Negative	Routine follow-up
2	Benign findings	Routine follow-up
3	Probably benign finding's	Short interval follow-up
4	Suspicious findings	Biopsy should be considered
5	Highly suggestive of malignancy	Biopsy required
6	Biopsy proven malignancy	–

1.3 Computer Aided Detection (CAD)

Automatic breast lesion identification in mammograms has been subject of study over the years. The objective is to assist the professionals in the screening tasks by automatically identifying microcalcifications and masses and reduce the number of FN and FP. To accomplish this task, several computer vision techniques where employed over the years, ranging from simple comparative methods to advanced machine learning techniques like Convolutional Neural Networks (CNN). Naive early approaches remount to the year of 1967 (Winsberg et al., 1967) where rectangular film segments containing breast image were collected and the optical distribution was carefully analyzed to providing pieces of evidence about the differences with normal images cases. Thresholding and fuzzy pyramid were employed by Brzakovic et al. (1990) to identify high-intensity homogeneous regions among mammograms. Next, the detected pixel groups are subjected to Bayes classification, considering the area, shape and edge features to classify between the benign and malignant cases. Many different image analysis approaches can be found. However is still unknown whether the use of Computer Aided Systems (CAD) methods yield better results than traditional diagnosis (Azavedo et al., 2012), as they still return a high number of false positives, diminishing specialists confidence in the system. Recent works (Technology, 2017) reported that CAD systems have surpassed radiologists. The development of more robust algorithms combined with the reduced cost of computational resources are the key ingredients to reduce the huge workload of mammogram screening and increase the accuracy of the diagnosis, leading to an increase of the survival rates. For this, becomes vital the combination traditional computer vision and deep learning techniques to obtain a robust system able to

handle large amounts of data while reducing the number of FN and FP diagnosis.

1.4 Objectives

This dissertation focuses on the development of a CAD system for breast cancer detections and diagnosis. The system encompasses several stages, ranging from image pre-processing to final image classification with the objective to aid clinicians in the task of screening and evaluating of mammograms images and reduce the percentage of human error that results in FN and FP diagnoses. Careful analysis of the methods employed on each of the stages provides insights about its robustness and potential impact in the complete CAD system. To construct a complete CAD system, the objectives of this dissertation relies on the accomplishment of the following steps.

- Deepening the knowledge in the field of computer vision for medical scenarios, specialty focusing on strategies that enable to overcome the imposed difficulties of low contrast lesion structures.
- Evaluation of the most promising state-of-the-art techniques and formulation of a robust architecture that allows a CAD system with high level of accuracy, to be applied to breast cancer images.
- Design and Implementation of a modern medium-sized architecture to fit nicely in a general purpose computer with limited resources, by exhaust evaluation and trials.

1.5 Contributions

The most important contributions of this thesis corresponds to the design of a modern medium-sized architecture for segmentation and classification tasks fits nicely general purpose computers where resources, data or computation power, is limited.

1.6 Outline of the Thesis

This thesis is structured as follows:

- Chapter 2 concisely presents literature review regarding breast region segmentation and classification.

- Chapter 3 lists the evaluated methods for breast structures segmentation providing pairwise comparison of the conducted experiments.
- Chapter 4 lists design and implementation details for the breast structures classification and results of the conducted experiments.
- Chapter 5 evaluates the CAD system as a whole identifying potential bottlenecks.
- Chapter 6 presents an overview of the accomplished results and pointing directions for future work.

Chapter 2

Literature Review

Pre-processing is the first step to be carried out in a computer vision system. In the case of mammograms images, typical pre-processing techniques can include: noise reduction (Romualdo et al., 2013), image enhancement (Wang et al., 2013), background exclusion, orientation homogenization (Li et al., 2013) and pectoral muscle identification (Akram et al., 2013) among others. The subsequent methods frequently benefit from the pre-processing stages, enabling them to focus and extract meaningful information for breast mammogram characterization and classification since redundant or non-relevant information has been properly processed. Ganesan et al. (2013b) states that the pectoral muscle segmentation methods can be categorized into different categories according to their main procedure,(Figure 2.1).

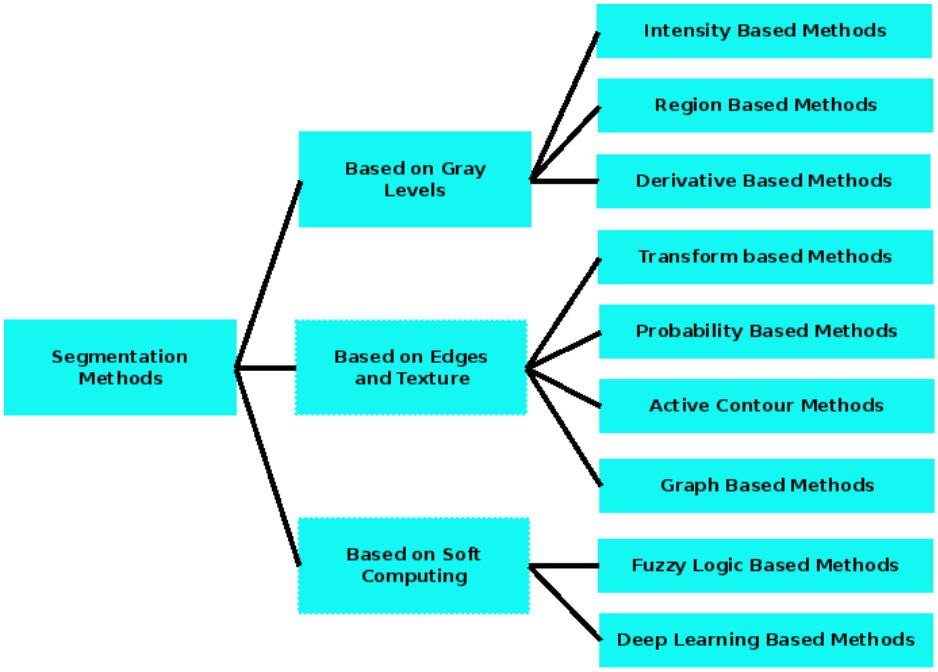


Figure 2.1: Mammogram main segmentation methods (Diagram based on Dey et al. (2016) survey)

2.1 Image Enhancement and Noise Reduction Approaches

Noise reduction (Romualdo et al., 2013) and image enhancement (Wang et al., 2013) are common primary tasks to be performed. For mammogram image enhancement, is common the use of Histogram Equalization or Contrast-limited adaptive histogram equalization (CLAHE) techniques (Reza, 2004). Filtering approaches are explored Gorgel et al. (2010) by the use of a wavelet transform of the mammogram image with the approximate coefficients being filtered by a homomorphic filter. The details coefficients associated with edges and noise are modeled by a Laplacian and Gaussian variables respectively. The obtained coefficients are compressed and enhanced by combining those variables with a shrinkage function. Finally, the fine details of the image are retained by using an adaptive threshold and the unwanted noise is suppressed. Morphological operations are commonly applied to mammogram images to remove artifacts like labels that might appear in an image. Those labels structures can be known apriori, and by applying morphological operations like hit and miss, artifacts can be removed from the image.

2.2 Pectoral Muscle Region Segmentation

After image enhancement and noise removal, is common to found a stage responsible for removal of the pectoral muscle tissue from mammogram image. The inclusion of the pectoral muscle on the image can bias the subsequent detection procedures, due to the fact that the majority of the methods to detect lesion masses relies on image intensity and correspondent density, that shares similarities with the pectoral muscle (Li et al., 2013; Akram et al., 2013).

Another important fact in pectoral muscle identification lies in the possibility that the local information may be contained among its edge. It's identification along with an internal analysis the region can be used to identify the presence of abnormal auxiliary lymph nodes, that are a common manifestation of an occult breast carcinoma (Ferrari et al., 2004).

The similarity of the pectoral muscle region with lesions, in the majority of cases, lead to the increase of the False Positives (FP) number and waste computational resources and time on image areas that are not relevant the for final diagnoses. For the pectoral muscle segmentation task, different approaches are presented by several authors and summarized in the following subsections.

2.2.1 Pectoral Muscle Region Segmentation Methods

For the task of pectoral muscle segmentation, several computer vision techniques can be employed to robust identify the region of interest.

Intensity-based segmentation is explored by Czaplicka et al. (2012) by the combination of multilevel OTSU Otsu (1979) to obtain multiple regions classes that are based on the number of gray levels to separate regions with a low-intensity background. The process is followed by a gradient estimation to produce a rough pectoral border, being smoothed by linear regression to attain the exact pectoral muscle contour.

Seed growing algorithms are explored by Maitra et al. (2012) to segment the pectoral muscle. The process starts by contrast enhancement of the image using CLAHE and later define a rectangular area to isolate pectoral muscle from the lesions Region of Interest (ROI) and finally suppressing the pectoral muscle by using a modified seed growing algorithm.

On the other hand, Molinara et al. (2013) employed the used of gradient-based methods. The process starts by employing a pre-processing step to normalize the image and highlight the pectoral muscle separating border. The gradient of the x axis among the highlighted image is considered by employing an edge detection followed by a Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981) algorithm to extract straight lines that separate the pectoral muscle from the neighbor regions. The algorithm exhibited good results in situations

were the pectoral muscle borders are nearly straight and strong, but the results were below average for images that presented curved pectoral border and faded border edges.

Local active contours are explored by Mencattini et al. (2012), that combined the use of local active contour scheme and Gabor filters, resembling to approach used by Ferrari et al. (2004). The processes start by decomposing the image using the Gabor Filters attaining magnitude and phase to create a 48 vector summation, used to detect the candidate line. However, in order to minimize miss candidates selection, the method starts by eliminating false pectoral edges candidates by applying different logical conditions, removing false candidate lines.

Liu et al. (2011) presented a statistical approach based on the idea of "Quality of Fit" to detect pectoral muscle edge. The method works on the basis of a joint normal distribution to determine the probability of a pixel belonging either a high or low-intensity region. Based on this decision, a contour is obtained to identify the pectoral muscle tissue. The algorithm assumes that the mammogram corresponds to a set of independent random intensity variables that can be modeled by a normal distribution.

Akram et al. (2013) proposed a pre-processing method to remove a pectoral muscle along with other artifacts. The method is based on the use of a modified active contour. The algorithm starts by thresholding the image using a $T = 15$, removing the low and high-intensity pixel labels along with scanning artifacts. Next, the pectoral border is traced using a multi-phase active contour and introducing of a new term M^k to the Mumford Shah model allowing to move the contour inwards and determining the stopping point from the difference between consecutive contour, deriving the final pectoral muscle contour.

Graph-based approaches usingShortest Path (SP) procedure were explored by Cardoso et al. (2010) to detect the pectoral muscle automatically. The process starts by transforming the image into polar coordinates and assuming the image new center of coordinates is located at the top left corner. Then a graph is constructed, were each pixel is a node connected to its neighbors by arcs. Each arc contains a weight value based on the gradient for that particular region. With the weighted graph formed, the optimal vertical paths are searched by employ a minimum cumulative cost C for each pixel nodes. Once the shortest path is constructed, the muscle edge is attained and the rows are transformed back to original Cartesian coordinate systems.

A variation of the SP technique is presented by Domingues et al. (2010), consisting of a two-step procedure to detect the muscle contour. In a first step, the endpoints of the contour are predicted with a pair of Support Vector Regression (SVR) models; one model trained to predict the intersection point of the contour with the top row while the other is designed for the prediction of the endpoint contour on the left column. Next, the muscle contour is computed as the SP in polar coordinates between the two endpoints. The input

features chosen to develop the models correspond to the gray-level values obtained from a 32×32 thumbnail of the cropped mammogram. Final result yielded Hausdorff Distance (HD) distance of 0.1232 on 150 mammograms from the INbreast database.

Novel supervised deep learning framework for region segmentation were proposed by Dubrovina et al. (2018) for region segmentation. The process aggregates regions into semantically coherent tissues using Convolutional Neural Networks (CNN) to learn discriminative features automatically. To overcome the difficulty involved with the used of a medium-size database, the training of the CNN was performed in an overlapping patch-wise manner. To accelerate the pixel wise-prediction, only the convolutional layers were used instead of the classical fully connected layers, enabling faster computations, while preserving the classification accuracy. The extracted patches were pre-processed prior to training to have zero mean and a loss accuracy multinomial logistic loss function was used. The results shown a Dice Coefficient (DC) of 0.85 regarding pectoral muscle region while the fibro-glandular tissue and nipple regions presented a DC equal to 0.61 and 0.56, respectively.

Petersen et al. (2014) presented a method to learn descriptive features from unlabeled mammograms. These learned features are used as the inputs to a simple classifier, addressing the following tasks: i) breast tissue segmentation ii) scoring of percentage mammography density (PMD), and iii) scoring of mammographic texture (MT). The employed texture scoring method learns a deep hierarchy of increasingly more abstract features from unlabeled data and maps the final feature representation to the label of interest. The pixel labels were grouped in the background (BG), pectoral muscle (PM), and breast tissue (BT). The training data was collected by randomly drawing 50,000 patches across a set of training mammograms associated with the true label and an unseen mammogram was segmented by applying the trained model in a sliding window approach. Results have shown a mean DC for automated vs. experts breast tissue segmentation of ($BG = 0.99, PM = 0.95$and $BT = 0.98$) regions.

Performance evaluation of methods for pectoral muscle segmentation is resumed in Table 2.1.

Table 2.1: Performance evaluation of pectoral muscle segmentation methods.

Year/Author	Main Method	# Images Success	Pros/Cons
Czaplicka et al. (2012)	Multilevel Otsu, Gradient estimation, Linear regression	300 MIAS, DICE 0.85	No wrong detection, Not robust
Maitra et al. (2012)	Contrast enhancement, Modified region growing	322 MIAS, CM 0.976	Simple, Not robust
Liu et al. (2012)	Iterative Otsu, Thresholding and morphological processing	150 MIAS, HD 0.087	Accurate, robust, Efficient, Computationally intensive
Akram et al. (2013)	Multi-phase active contour	MIAS, DICE 0.771, Sens 0.978	Accurate with good pre-processing, Not robust
Cardoso et al. (2010)	Polar Transformation and Shortest path	INbreast, HD 0.86	Simple, efficient
Domingues et al. (2010)	Endpoints using SVR SP in polar coordinates	150 INbreast, HD 0.1232	Simple, efficient
Dubrovina et al. (2018)	Semantic Segmentation CNN and patches	40 Images, DICE 0.85	Works well on muscle border near dense tissues
Petersen et al. (2014)	Semantic Segmentation CNN and patches and sliding window	50,000 patches, DICE 0.95	Robust, Multiclass segmentation

FP (False Positive - lower the better), Acc (Accuracy - higher better), HD (Hausdorff Distance - lower the better), DICE (higher better), CM (Combined Measure), Measures range $[0, 1]$

2.2.2 Summary

The overview of the different techniques covered in this section focuses on the efforts made towards solving the pectoral muscle segmentation problem in the pre-processing stage of the Computer Aided Systems (CAD) system. The discussion about the different methods proposed by researchers among the literature reveals that very few methods can achieve accurate results on a wide range of images with varying position, shape and size of the pectoral muscle. The performance of the enlisted methods is useful for comparison purpose, giving insights on how to devise a robust, yet simple pectoral muscle extraction algorithm that achieves a high accuracy.

2.3 Detection and Characterization of Suspicious Regions

Having segmented pectoral muscle and screened out normal mammograms, the following task typically involves looking for suspicious regions in mammogram images. Two types of findings can be found, calcification's and masses. Due to their differences, specialized detection and characterization methods are employed on each of the cases.

2.3.1 Mass Lesions Detection

Several computer vision techniques can be employed for the automatic detection of masses. A summary of those techniques is presented in Figure 2.2

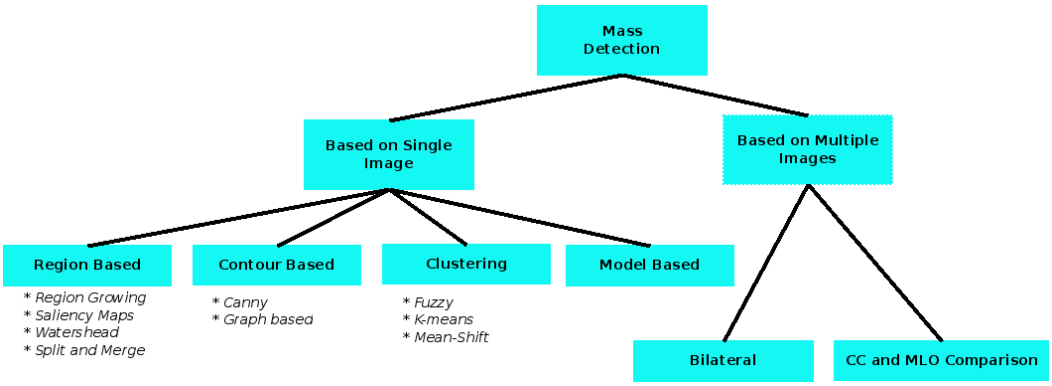


Figure 2.2: Hierarchy of mass detection main methods.

Multi-view mammographic analysis, with the main focus on breast cancer detection at a patient level was explored by Velikova et al. (2009). The main objective of the multi-view detection is to determine whether or not the object has certain characteristics (e.g., being cancerous) by establishing correspondences between the 2D image characteristics of regions (subparts) in multiple object views (projections). The modeling scheme is based on two Bayesian networks with a hand-constructed (fixed) structure to explicitly represent the multi-view dependences in the detection problem, enabling that the two different regions A_i and B_j that are generally conditionally independent become dependent once exist evidence that they are the projections of the same lesion in two views.

The multi-view scheme was also proposed by Ericeira et al. (2013) making use of bilateral information. Asymmetric regions regarding the left and right breast mammograms pair are detected by means of structural variations between corresponding regions, making use of a spatial descriptor defined as a cross-variogram function. After determining the asymmetric regions, the variogram function is applied to each asymmetric detected region separately in order to be classified as either mass or non-mass. Results on the Digital Database for Screening Mammography (DDSM) database were 0.9026 of accuracy, 1.00 sensitivity and

0.8537 for specificity.

Agrawal et al. (2014) used saliency-based segmentation to obtain the mass regions. The Graph-Based Visual Saliency starts by computing the saliency of a region with respect to its local neighborhood by exploring the use of directional contrast. Three main steps are involved during this task: (1) Feature maps computation from contrast maps over four Gabor filter orientations $[0°, 45°, 90°$ and $135°]$; (2) activation's and normalization of maps; (3) combinations of the normalized activation's. The final region segmentation is obtained by thresholding the saliency map. From those segmented regions, several features are extracted to serve as input to a Support Vector Machine (SVM) classifier for mass classification. Experiments were carried out using the Mammographic Image Analysis Society (MIAS) database. The results have shown that 82% of the 58 masses were detected and each image presented 2 to 3 FP detections.

Pereira et al. (2014) explores the use of Ipsilateral information. The process starts with a pre-processing the image using wavelet decomposition and Wiener filtering for image de-noising and enhancement. The segmentation of suspicious zones is achieved using a Genetic algorithm. A manual post-processing step is carried out in particular areas were marked structures from the Craniocaudal Mammogram (CC) view are compared with the Mediolateral Oblique Mammogram (MLO) view. The results shown a FP rate of 1.35 FP per image with a sensitivity of 0.95 using DDSM database.

A summary of the described methods and the performance is presented in Table 2.2.

Table 2.2: Performance evaluation of mass detection methods.

Year/Author	Main Method	# Images Success	Pros/Cons
Velikova et al. (2009)	Multiple view fusion using Bayesian networks	1063 , AUC 0.862	Comparison two images
Ericeira et al. (2013)	Asymmetric regions, cross-variogram spatial descriptor	Acc 0.903, Sens 1.00	Comparison two images
Agrawal et al. (2014)	Graph Based Visual Saliency	58 MIAS, Acc 0.82	Efficient
Pereira et al. (2014)	Pre-processing, Ipsilateral information	MIAS, Sens 0.95	Manual post-processing

Acc (Accuracy - higher the better), Sens (Sensibility - higher the better), AUC (Area Under the ROC Curve - higher the better) Measures range $[0, 1]$

2.3.2 Mass Lesion Contour Extraction

For mass contour extraction, several methods can be employed to obtain the lesion contour. Several methods can be employed for the task.

Domínguez and Nandi (2009) described a Dynamic Programming (DP) technique for segmenting medical regions. The method starts by constructing a local cost function, assigning a cost to each pixel in a polar representation to obtain a cumulative cost matrix. The contour lesion is defined by those pixels that linked together form a path with the lower cumulative cost. A particular DP-based segmentation algorithm is used, ID^2PBT that can be described in three main components: (1) Edge strength component were pixels with strong edge content are assigned to a low cost, and vice-versa; (2) Gray-level component in which pixels gray-level values are similar to a preferred gray level (which corresponds to the boundary of masses), are also assigned to a low cost; (3) Shape component where the shape of each particular mass to be segmented is modeled by an ellipse. The ellipse parameters (axes and orientation) are computed by the algorithm based on the initial estimative of the boundary of the mass. Finally, the three components are linear combined, creating total cost function, and the weights of each linear component are dynamically adjusted by the ID^2PBT algorithm, based on the relative agreement of the components. The ID^2PBT results were compared with customized region growing (CRG) segmentation methods. The average Area Overlap Measure (AOM) between Ground Truth (GT) and the set of contours of each segmentation method was 0.72 and 0.83 respectively among 348 masses.

Rabottino et al. (2008) used region growing technique to select ROI candidate spots to extract shape and texture features supplied to a fuzzy classifier. Extensive experiments were conducted using the DDSM database. The result shown a CM (Completeness) of 0.8834 and CR (Correctness / TP rate) of 0.9338.

Tizhoosh et al. (2016) presented a segmentation method based on Content barcodes. A binary descriptor based on Radon transform is used to find similar cases and estimate the surrounding tumor bounding box. The approach starts by indexing all available GT images by first assigning two barcodes for each bounding box (of each ground-truth): a "global" barcode for the entire image, and a "local" (ROI-based) barcode for the obtained bounding box. ROI estimation was subsequently performed through a search of similar cases on the database. When querying a new image, a fixed-size ROI is first defined by asking the user to provide a mouse click in the center of a tumor. Then the query image is subsequently tagged with two barcodes (global and ROI based). Using a similarity measure the barcodes obtained from the query image are compared with barcodes from images in the training set. This enabled to identify the top most similar tumors and estimate the location of a tumor in the query image. Experiments with 33 B-scan images resulted in promising results, exhibiting an accuracy of 0.81.

Cardoso et al. (2015) explored the use of SP algorithms to obtain the mass contour. The computation of the closed contour is accomplished in the original coordinate space instead of transforming the image into polar coordinates. After defining a directed acyclic graph, one of the main difficulties in operating in the original coordinate space is addressed by modulating the cost of the edges to counterbalance the bias introduced by the small paths that collapse in the seed point. The first task involves creating a Directed Acyclic Graph (DAG) from the grid with a proper linearization while the second step addresses paths closer that are closer to the center of the region that contains fewer pixels, avoiding being selected. Third, a Euclidean distance between nodes (pixels) is defined to capture the distance in the context of closed paths that enclose a given node. Experiments were conducted in INbreast database and several results using different seed locations and perturbation were obtained. The method exhibit a AOM of 0.788 for a *alpha* of 2.

A summary of the contour extraction methods is presented in Table 2.3.

Table 2.3: Performance evaluation of mass contour extraction methods.

Year/Author	Main Method	# Images Success	Pros/Cons
Domínguez and Nandi (2009)	Dynamic Programming, polar representation	Private DB, Acc 0.72	Simple, Polar distortion
Rabottino et al. (2008)	Region Growing	-, CM 0.8834	Simple, Not robust
Tizhoosh et al. (2016)	Content barcodes, SVM	33 B-scan, Acc 0.81	Simple, Manual input
Cardoso et al. (2015)	Shortest path algorithms, Original Coords	INbreast, AOM 0.788	Original Coordinates, Robust

Acc (Accuracy - higher the better), CM (Combined Measure - higher the better), AOM (Area Overlap Measure - higher the better), Measures range $[0, 1]$

2.3.3 Calcification Lesion Detection

The methods for calcification detection can be grouped mainly into four categories, (1) simple image enhancement methods, (2) multi-scale decomposition, (3) stochastic modeling and (4) machine learning methods, or combinations of the mentioned methods.

Top-hat transformation combined with wavelet decomposition for image enhancement and de-noising respectively was purposed by Zhang et al. (2013). The calcifications are detected based on the feature distribution. Results showed that 92.9% of the true calcification's were detected presenting an average of 0.08% FP per image.

Huang et al. (2013) uses also top-hat filtering and wavelet transformation for calcification

detection. Top-hat address the uneven background while wavelet extracts high-frequency components from the image. Additionally, the use of Laws filter allows further feature extraction. The final set of candidates is constructed in an interactive method by performing morphological and edge detection followed by SVM based classifier to reduce the number of FP. Results shown a sensitivity of 92% and a Area Under ROC curve (AUC) of 0.99 and 0.65 FP per image.

Zhang et al. (2014) combines the morphological methods with a SVM classifier. Initially, the image is subject to contrast correction and two structural elements are employed to enhance potential calcification's. Next, dual threshold extracts potential regions and the SVM classifer is used to reduce the number of FPs. The experiments were conducted on the MIAS database, achieving a True Positives (TP) rate of 0.9885, a FP rate of 0.782, presenting 0.53 FP calcification's per normal mammogram.

Deep learning approaches for calcification are explored by Shin et al. (2014). Local peak detection scheme attains potential calcification's regions. Patches of the detected areas are then manually annotated as calcification's or not. A Discriminative Restricted Boltzmann Machine is applied to automatically learn calcification's morphology and the obtained model is used to classify new images patches. Results shown a AUC of 0.903 using 322 images from MIAS database and 280 from privately owned database.

Wang et al. (2016) evaluated the performance of deep-learning methods for micro-calcification detection and classification. Segmentation is performed by a semi-automated method to characterize all micro-calcifications. To asses the quality of the model, several features are extracted to evaluate the performance of traditional methods. SVM, Linear Discriminant Analysis (LDA) and k-Nearest Neighbours (kNN) serve as benchmark the models for the deep learning model. The deep learning model achieved a discriminative accuracy of 0.873 if micro-calcifications were characterized alone, compared to 0.858 with a support vector machine. The accuracy was 0.613 for both methods when only masses were present, being improved up to 0.897 and 0.858 after micro-calcifications combined analysis. Overall, deep learning models obtained from large datasets presented superior performance when compared to standard methods for the discrimination of micro-calcifications. Accuracy increased by adopting a combinatorial approach to detect microcalcifications and masses simultaneously.

A summary of the described methods, namely the main methodology, databases, performance and pros/cons is presented in Table 2.4.

Table 2.4: Performance evaluation of micro-calcification detection/classification methods.

Year/Author	Main Method	# Images Success	Pros/Cons
Zhang et al. (2013)	Top-hat filtering and wavelet transformation, Laws filter	Private DB, Sens 0.929	Simple, High FP
Huang et al. (2013)	Top-hat filtering and wavelet transformation, SVM	MIAS Sens 0.985	Simple, High FP
Zhang et al. (2014)	Morphological methods, SVM	MIAS, Priv DB, Sens 0.988	Simple, High FP
Shin et al. (2014)	Local peak detection, Discriminative Restricted Boltzmann Machine	322 MIAS, 280 Private, ROC 0.902	Simple, Manual annotations
Wang et al. (2016)	Deep Learning, Combined Calcification's and Masses	1204 Private DB, Acc 0.873, Combi Acc 0.897	Simple, Manual annotations

Acc (Accuracy - higher the better), Sens (Sensibility - higher the better), ROC (Receiver operating characteristic - higher the better), Combi Acc (Combined Accuracy - Masses and Calcifications), Measures range $[0, 1]$

2.3.4 Summary

The overview of the different techniques covered in this section focuses on the efforts made towards solving the detection of lesions that appear in mammogram images. The discussion about different methods proposed for detection and characterization of the findings allows defining two main groups of techniques, one for masses and other for calcifications. Several methods that range from simple morphological operations to deep learning approaches in more recent years are used to detect lesion findings. Since the external contour is used to characterize the malignancy of the findings, it's important that methods extract with great accuracy the lesion boundary. The non-uniform contrast nature of the findings adds additional difficulties for contour extraction.

2.4 Classification of Breast Images and Lesions Findings

The process of classification of breast image is commonly focused in the evaluation of the masses and calcification lesions by CAD systems. It consists of the computation of numerical values to quantify certain object or region properties. The Breast Imaging Reporting And Data System (BI-RADS) standard recommends the description of calcification's according to their distribution and morphology, while masses are mainly characterized through their margins, shape and density characteristics.

2.4.1 Mass Lesion Classification

For mass classification several features are commonly extracted to characterize the lesions, being useful to construct computational models enabling to classify new unlabeled mammogram images. Recent techniques are fusing these two concepts (feature + classification) into a single one, with features being automatically learned by the models instead of being defined apriory to the classification task. Mass features can range from simple texture characterization up to detailed contour description to capture mass characteristics.

Rangayyan et al. (2000) focus the analysis in less common shapes of masses, namely circumscribed malignant tumors and spiculated shape benign masses that are difficult to classify. The proposed method relies on segmentation to separate major portions of the boundary and label them as concave or convex segments. Over these segments, features are extracted that characterize the concavity fraction and degree of narrowness of spiculated index.

Sampat et al. (2005a) employed the use of a Beamlet transformation to characterize lesions into 4 main categories: round, oval, lobulated, or irregular. A kNN is then employed for classification. Using the DDSM database, a set of 25 images of each type was used to test the method, obtaining an accuracy of 0.78 for classifying masses as oval or round and an accuracy of 0.72 for lobulated or round masses.

A new mass descriptor is proposed by Cheikhrouhou et al. (2008). Its based on geometrical feature, perimeter, and three morphological features (contour derivative variation, skeleton endpoints and spicularity). The descriptors were evaluated on DDSM using SVM classifier fitted with a Gaussian kernel . It achieved an accuracy of 0.93 for the two class case (malignant and benign) and 0.857 for the four class model (BI-RADS I, II, III and IV).

Rojas-Domínguez and Nandi (2009) presented four new features designed to be invariant to the exact shape of the contour. The first feature quantifies the degree of spiculation of a mass and its likelihood of being spiculated, while the second measures the amount of mutual information between selected components of the mammography images. The remaining two features measures the local fuzziness of the mass margins in automatically selected points. All those features characterize the (circumscribed/spiculated) of the masses, enabling to identify (benign/malignant) masses cases that occur in MIAS and DDSM databases. A SVM model was then trained, and in combination with the computed features, exhibited a 0.89 correct classification. In BI-RADS diagnosis, the the performance was approximately 0.81 for correct classification.

Two new shape measures for quantifying the degree of convexity are proposed by Rosin (2009). The first is based on convexification, while the second in contained lines. The experiments were conducted on a set of 54 masses from mammograms from MIAS and

private databases. kNN was the selected classifier using the Mahalanobis distance. A correct classification of 0.944 was achieved on circumscribed/spiculated discrimination, 0.741 for benign/malignant discrimination, 0.684 for circumscribed benign/circumscribed malignant/spiculated and benign/spiculated malignant discrimination.

Content-based mammogram retrieval system was proposed by Wei et al. (2011) making use of three mass features to describe shape (Zernike moments), margin (sharpness degree), and density (density degree).

To circumvent the problem of non-invariance to the rotation for round-shaped objects, Moura and López (2013) proposed a new descriptor, the Histograms of Gradient Divergence (HGD), to address the problematic. The method quantifies the gradient angle divergence towards the center of the lesion. The feature capabilities were compared with 11 conventional image descriptors applied to DDSM and Breast Cancer Data Repository (BCDR) databases using a SVM classifier. Overall, HGD scored best on both databases when classifying masses as benign or malignant.

Vadivel and Surendiran (2013) presented new geometric shape and margin features to characterize mammogram mass lesions into 4 categories: round, oval, lobular and irregular. Experiments were conducted on mammogram images from DDSM database in combination with a C5.0 decision tree classifier. It yielded an accuracy of 0.8776 for irregular, lobular, oval or round cases, 1.00 for oval versus round, and 0.9545 for lobulated vs round.

Liu and Tang (2014) describes a mass classification system that encompasses the use of geometry and texture features combined with a SVM-based feature selection method. After segmentation, a set of geometric and texture features are extracted by taking advantage of the fact that typical benign mass presents a round, smooth and well-circumscribed boundary, whereas the boundary of a malignant tumor is usually spiculated, rough, and blurry (Mudigonda et al., 2000). After boundary analysis, the extracted geometric features characterize the shape of the mass boundary contour. To access the quality of the method, compactness (Kilday et al., 1993) is used to measure the level of complexity vs the enclosed area.

Tan et al. (2014) used 181 image features that describe the mass shape, spiculation, contrast, presence of fat or calcifications, texture, isodensity, and many other morphological characteristics. For feature selection, a sequential forward floating feature selection-based method was applied. The system performance was assessed using a SVM classification model applied to 1200 ROI's images (600 malignant masses and 600 benign), randomly selected from a private and DDSM databases. It yielded a AUC of 0.805. The more relevant features were those related to mass shape, isodensity and presence of fat, which is consistent with the image features that radiologists frequently use for supporting their mass classification decisions.

To address the problem of the selecting the most informative features for modeling, Pérez (2015) presented a new feature selection method named $uFilter$ based on the Wilcoxon rank sum, (McKnight and Najab, 2010) to ranking relevant features, which asses the relevance of features by computing the separability between the class-data distribution of each feature. The $uFilter$ method effective ranks relevant features independently of the samples sizes, making tolerant to unbalanced training data, does not require any type of data normalization and reduces the risk of data overfitting.

Multiple Kernel Learning (MKL) approaches were explored by Santos (2017) enabling the creation of models where each feature can be treated in a different way, that may improve the quality of the learned models. The imbalance problematic was tackled by adopting a strategy of weighing the benign and malignant cases in order to produce models that are more reliable and robust to the class distribution. Results show that the weighted approach produces better quality models for both balanced and unbalanced mammogram datasets when using MKL approaches.

A brief summary of the features that can be used to characterize mass lesions is presented in Table 2.5.

Table 2.5: Summary of the features used in the literature for mass characterization.

Type	Feature	Short Acronym
Contour	Acutance Histogram (Tao et al., 2008)	AcH
	Circularity (Vadivel and Surendiran, 2013)	Circ
	Concavity fraction (Rangayyan et al., 2000)	fCC
	Contained lines (Rosin, 2009)	Cl
	Curvelets (Moura and López, 2013)	Curv
	Eccentricity (Vadivel and Surendiran, 2013)	ECT
	Elongatedness (Vadivel and Surendiran, 2013)	En
	Extent	Ext
	Fuzziness of mass margins (Rojas-Domínguez and Nandi, 2009)	FZk
	Major Axis Length	MJl
	Maximum radius (Vadivel and Surendiran, 2013)	Rmax
	Minimum radius (Vadivel and Surendiran, 2013)	Rmin
	Radial to tangential signature (Rojas-Domínguez and Nandi, 2009)	SpSI
	Perimeter (Vadivel and Surendiran, 2013)	Per
	Shape Index (Vadivel and Surendiran, 2013)	ShI
	Sharpness (Wei et al., 2011)	Sh
	Skeleton end points (Cheikhrouhou et al., 2008)	SEP
	Spiculation (Cheikhrouhou et al., 2008)	Sp
Texture	Gabor filter banks (Moura and López, 2013), Saranya and Samundeeswari (2016)	Gab
	Grey-level difference matrix (Moura and López, 2013)	GLDM
	Grey-level run length (Moura and López, 2013)	GLRL
	Haralick (Haralick et al., 1973)	HaR
	Histograms of oriented gradient (Moura and López, 2013)	HOG
	Wavelets (Moura and López, 2013)	Wav
Statistical	Area (Vadivel and Surendiran, 2013)	A
	Beamlet (Sampat et al., 2005b)	Beam
	Compactness (Vadivel and Surendiran, 2013)	Com
	Convexification (Rosin, 2009)	Cvf
	Curvature Scale Space (Tao et al., 2008)	CSSD
	Energy	Ener
	Equivalent diameter (Vadivel and Surendiran, 2013)	Eqd
	Fourier (Tao et al., 2008)	NFD
	Dispersion (Vadivel and Surendiran, 2013)	Dp
	Entropy (Vadivel and Surendiran, 2013)	Entpy
	Mass edge Std (Vadivel and Surendiran, 2013)	Esd
	Mass Std (Vadivel and Surendiran, 2013)	SD
	Spiculation Index (Rangayyan et al., 2000)	SpI
	Histograms of Gradient Divergence Moura and López (2013)	HGD
	Contour Derivative Variation (Cheikhrouhou et al., 2008)	CDV
	Relative gradient orientation spiculation (Rojas-Domínguez and Nandi, 2009)	SpGO
	Zernike Moments (Moura and López, 2013)	Zm
	Convexity fraction (Rangayyan et al., 2000)	fCV
	Euler number (Vadivel and Surendiran, 2013)	EULN
	Texton (Tao et al., 2008)	Txo
	Thinness ratio (Vadivel and Surendiran, 2013)	Thi

2.4.2 Breast Image Classification using CNN

Recent developments in image classification, namely CNN models have been adopted by the scientific community to address the problematic of mammogram segmentation and classification. CNN encompasses in a single task the problematic of feature extraction and classification.

Shen (2017) developed an end-to-end training algorithm to classify whole image breast mammograms. Annotated lesions were required only at first stages of training. A CNN was recursively defined to obtain a patch classifier, and by simply adding a new convolutional layer on top of the final trained patch classifier, the system was transformed into a whole image classifier by modifying the input layer. This enables the finely tuned patch classifier to be efficiently used as a scan method of the whole image in one single forward propagation, generating predictions for the overlapping patches of the image and generating a heat map (representation of the likelihood of each patch belonging to one of the class patches). On DDSM, the model achieved a per-image AUC score of 0.88 and three-model average increased the score up to 0.91.

Multiple models were proposed by Dhungel et al. (2017) to detect, segment and classify individual mammogram images. For the detection's stage, the author uses a cascade of simple to complex classifiers to screen out obvious negatives cases and attain only the positive cases to be processed by the next stage. A sequence of three Deep Belief Network (DBn) models were trained, with the first stage responsible for the classification of pixels into positive (belonging to lesion) or negative (not belonging to lesion), starting from a coarser resolution towards a finer resolution in the last stage of the cascade classifier. The obtained positive pixels are combined by union with a Gaussian Mixture Model trained only at the finer resolution. After this, connected pixels are considered as potential lesions and proceed to the next stage, consisting by two sequential CNN models that follow the same cascade approach, with the second model only seeing previously positive classified examples. A final third stage is employed, consisting in two sequential Random Forest (RF) classifiers trained on a set of handcrafted features. The system detected 90% of masses at 1 per image, obtained a segmentation accuracy around 0.85 (Dice Coefficient) on the correctly detected masses, while the classification as malignant or benign presented a sensitivity of (Se) of 0.98 and specificity (Sp) of 0.7.

Carneiro et al. (2015), makes use of a full mammogram images to train CNN models. The process starts by resizing images to 264×264. The deep model receives as input one image corresponding to a mammogram and two binary images corresponding to each lesion segmentation per CC and MLO views and corresponding mass and calcification's segmentation of the same breast, forming an individual CNN. Each segmentation map uses an Imagenet pre-trained model, integrating the information at a later layer in the network.

To keep details of the diagnosis, bigger models were required at the cost of more examples to train. Results yielded a AUC (for a 2-class problem - benign and malignant) over 0.9.

A summary of the deep learning approaches for mammogram classification is presented in Table 2.6.

Table 2.6: Performance evaluation of deep leaning methods.

Year/Author	Main Method	# Images Success	Pros/Cons
Shen (2017)	Patch classifier, whole classifier, models combinations	2584 DDSM, AUC 0.91	Good results, Memory intensive, Generalization to other databases
Dhungel et al. (2017)	Cascade classifier, Cascade CNN , Detect, segment and classify	INbreast, Sens 0.98, Spe 0.7	Good results, Memory intensive, Complete pipeline
Carneiro et al. (2015)	Full image CNN and Imagenet pre-trained segmentation map	INBreast and DDSM, AUC $[0.9, 0.95]$	Memory intensive, Can overfit, Required large data
Sens (Sensibility - higher the better), Spe (Specificity - higher the better), AUC (Area under the curve - higher the better). Measures range $[0, 1]$			

2.4.3 Summary

This section focuses on the different techniques to asses the severity of the findings, by employing binary or multi-class classifiers. traditional computer vision methods commonly encompass a set of extracted features that attain properties of the findings that are associated with is the level of malignancy. Common features range form contour description, textural and statistical proprieties used to train models, that in the presence of new unseen examples extract the same evaluated features and predict the severity of the finding. Recent with deep learning techniques have emerged in the context of medical imaging, taking advantage of the fact that the features and classifier are combined into a single entity to be optimized. This simplifies the task of finding the best-describing feature since this can be learned during training. The major disadvantage of these techniques is the need of large labeled datasets, that in the context of medical imaging is still a limitation.

Chapter 3

Breast Structures Segmentation

This chapter corresponds to the segmentation stage of the Computer Aided Systems (CAD) system. Section 3.2 describes both INbreast (Moreira et al., 2012) and the Breast Cancer Data Repository (BCDR-D01) (Lopez et al., 2012) databases that correspond to the same patients, containing the characterization of the findings and corresponding annotations. The segmentation stage is divided into two main components: First, (Section 3.3) corresponds to the initial stage of the CAD system, detailing the pre-processing methods responsible for enhancing images and segmenting pectoral muscle. Second, (Section 3.5) provides details about the methods employed for lesion detection and contour extraction. Conducted experiments and results are presented in the final of each of the sections.

Lesion classification is detailed in Chapter (4).

3.1 Solution Overview

We focus our efforts on the development several stages of a CAD system to fulfill the objective of a complete solution for breast image diagnosis. Each of the stages of the CAD system,

are carefully compared with state of the art procedures, ranging from segmentation, contour extraction, feature extraction and classification, summarized in Figure 3.1.

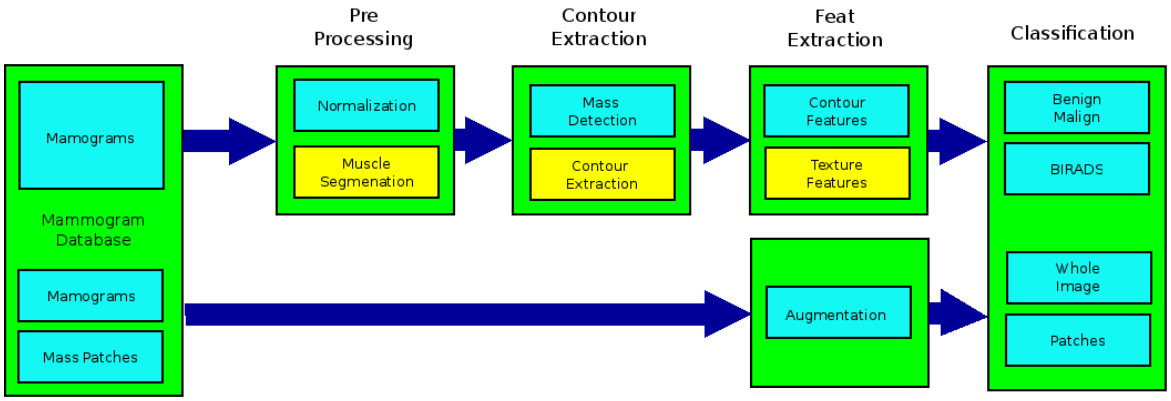


Figure 3.1: Frameworks main components.

The first stage, described on current Chapter (3), focuses in the segmentation stage that encompasses pectoral muscle and lesion segmentation. For the pectoral muscle segmentation, Shortest Path (SP), Region Growing, Intensity-based, Active contour and semantic segmentation using Encoder-Decoder architecture are carefully evaluated and results are presented. For lesions detection and subsequent characterization, saliency maps, watershed and Iris filter methods are used followed by a False Positives (FP) reduction stage. This enables to obtain the locations of potential lesions areas, to later characterize its exact contour.

The second part of the system, Chapter (4), focuses on the classification of the characterized lesions to asses its malignancy. Two types of classification task were conducted, a binary benign/malign classification and Breast Imaging Reporting And Data System (BI-RADS) ranking using re-sampling and ordinal models. In addition, Convolutional Neural Networks (CNN) were used to detect lesions and to classify its findings.

3.2 INbreast Database and Findings Description

INbreast (Moreira et al., 2012) images were acquired at a breast center located in a university hospital (Centro Hospitalar de S. João [CHSJ], Breast Centre, Porto). MammoNovation Siemens full-field digital mammography (FFDM), with a solid-state detector of amorphous selenium was used to obtain the images. The INbreast database contains a total of 115 cases (410 images) from which 90 cases are from women with both breasts affected (four images per case) and 25 cases from mastectomies (two images per case), with findings distribution described in Chart 3.2.

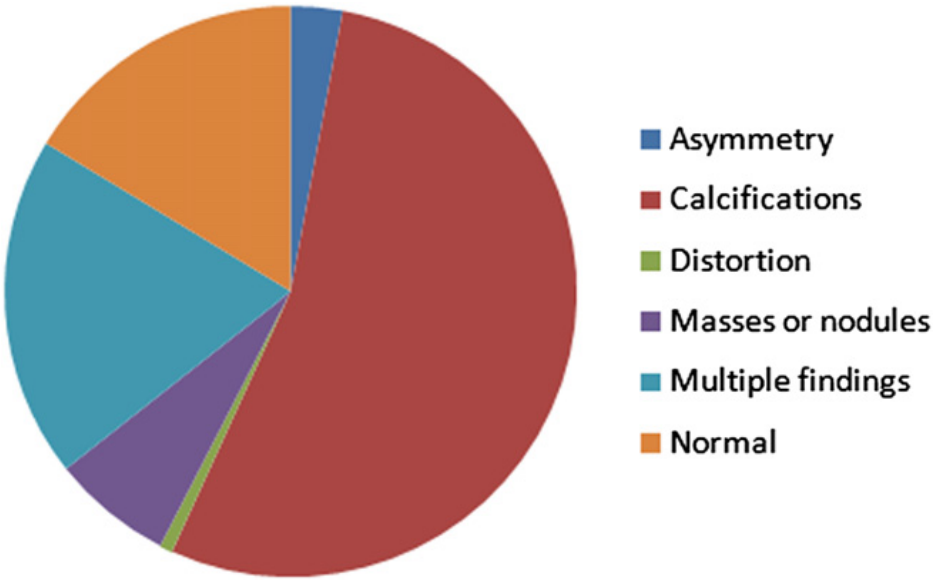


Figure 3.2: Chart describing the findings in the INbreast database (Image from Moreira et al. (2012)).

According to BI-RADS, a mass is defined as a three-dimensional structure exhibiting convex outward borders, usually evident on two orthogonal views. Benign calcifications are usually larger than calcifications associated with malignancy, often round with smooth margins and easily visible in images. Calcifications associated with malignancy are usually very small. Architectural distortions are defined as a focal interruption of the normal mammographic pattern of lines (converging at the nipple), usually presenting a star-shaped distortion, with no definite mass visible. Asymmetry on the other hand lacks convex outward borders of a mass and can be represented in three ways: size asymmetry (difference in the volume between the right and left breast), focal asymmetry (a unilateral, localized area of parenchyma), and global asymmetry (difference in the amount of parenchyma between the right and left breast) (D'Orsi, 2013).

The Ground Truth (GT) annotations were made by a specialist in the field, and validated by a second specialist. Each of the findings has an associated label that identifies the type of lesion. Several examples of GT annotations can be found in the database (Figure 3.3).

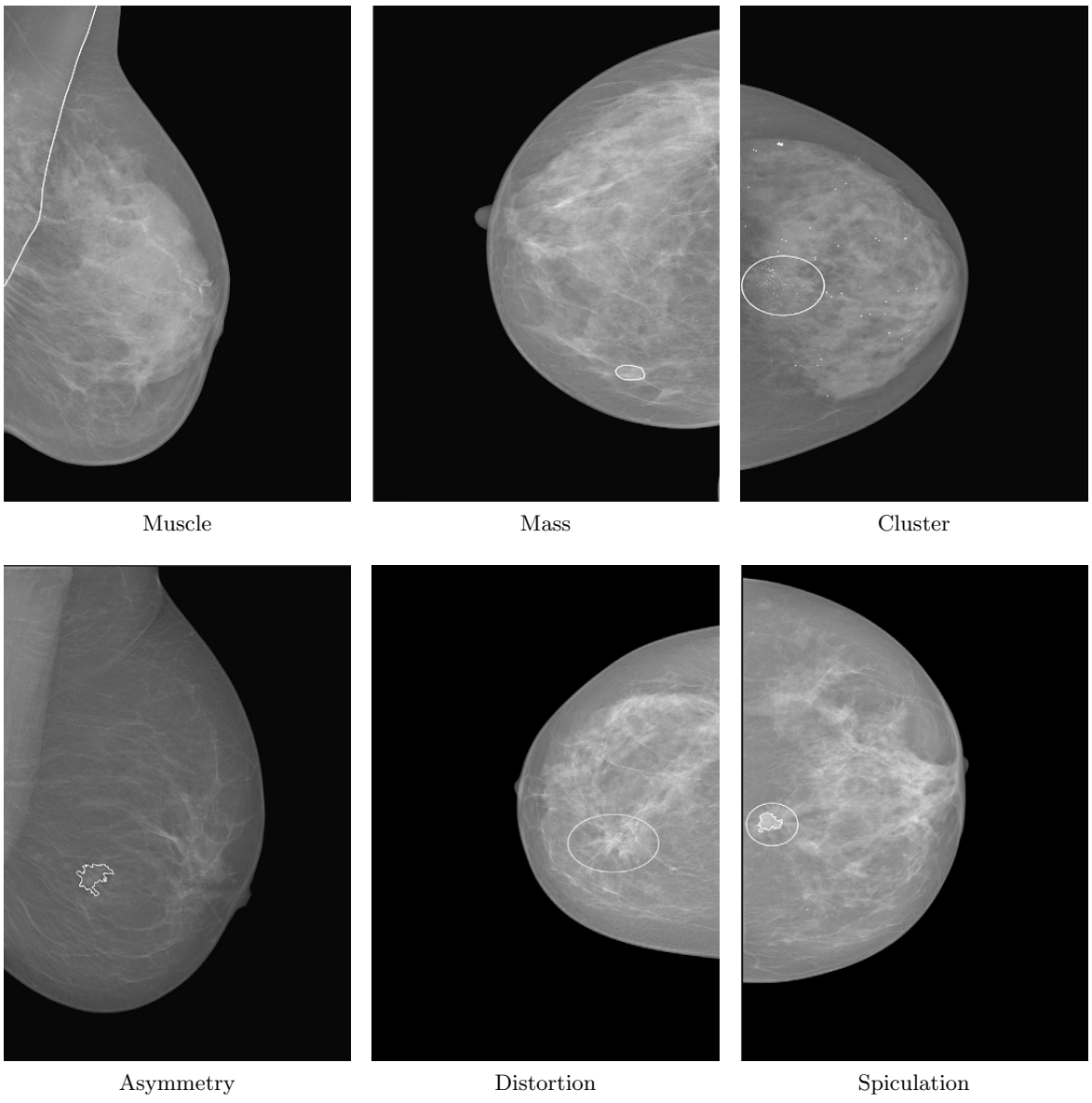


Figure 3.3: Example of the GT annotations.

3.3 Image Enhancement

Image enhancement is the first step to be carried out in a computer vision system by enhancing or removing image structures, increasing the chance of success of the subsequent algorithms. The most common tasks are contrast enhancement and artifact removal. Images denouncing low contrast or artifacts can degrade to a great extent the performance of the CAD system. Contrast enhancement is responsible for increase image details by mapping pixel intensities to a uniform distribution, while morphological operations are commonly employed to remove artifacts that can be present on images and enhance specific image structures.

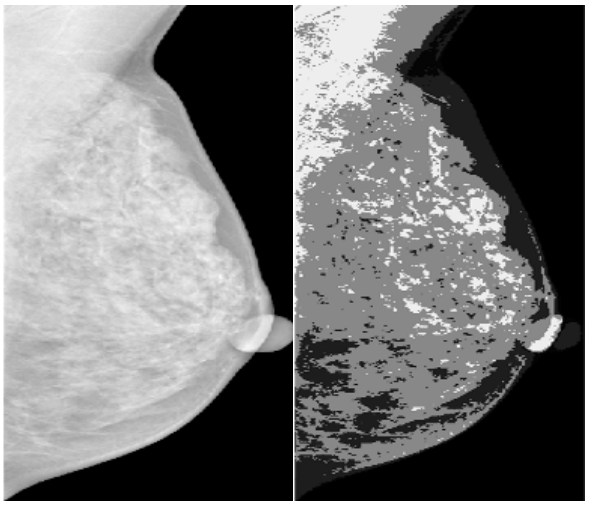
3.3.1 Image Normalization and Histogram Equalization

Image Normalization acts in the global image domain to normalize the image range between $[0, 1]$ by computing a new pixel intensity $I_{(x,y)}$ as

$$I_{(x,y)} = \frac{I_{(x,y)} - \min(I)}{\max(I) - \min(I)}. \tag{3.1}$$

.

Histogram equalization (Kim, 1997) is a technique for adjusting image intensities to enhance contrast. The method is useful in images with backgrounds and foregrounds that are both bright or both dark (Figure 3.4).

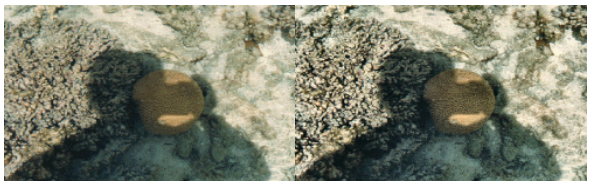


Original Equalized

Figure 3.4: Original image on the left and modified image on the right.

3.3.2 Contrast-Limited Adaptive Histogram Equalization (CLAHE)

Extensions to the histogram equalization were developed to address non-uniform image illumination. Contrast-limited adaptive histogram equalization (CLAHE) Haralick et al. (1987); Zuiderveld (1994); Lee et al. (2015) starts by constructing local histograms that encompass several square regions in the surroundings of a given point and employing a histogram equalization for each of the areas. Figure 3.5, exemplifies the use of a predefined image block size to compute local contrast values around the pixel $I_{(x,y)}$ were histogram equalization must be performed. The final pixel values are interpolated from the four closest local histograms.



Original CLAHE

Figure 3.5: Contrast enhanced image comparison.

Block size must be larger than the feature to be preserved, and the number of bins is directly related to the selected block size. Limits to the contrast stretching are defined by setting a max slope value to the intensity transfer function. CLAHE also prevents the over-amplification of noise on relatively homogeneous regions by establishing a maximum number of pixels that can have the same intensity, redistributing those uniformly for each local histogram grid.

3.3.3 Morphological Operations

Morphological operations (Haralick et al., 1987) are a set of non-linear filters used to process objects in the input image based on their shape, encoded by a structuring element. Simple bit-wise operations like *Union, Inversion, Intersection* or combinations can be performed between the structural element and the input image.

Dilation $(I \oplus S)$ denotes binary dilation between the image I and structure element S as

$$I \oplus S = \{(p+q)|p \in I, q \in S\}. \tag{3.2}$$

The dilation operation is obtained by translating a point p in the image with a point q in the structured element. Basically encompasses the union of the structuring element S_p copies, centered at every pixel location p on the foreground

$$I \oplus S = \bigcup_{p \in I} S_p. \tag{3.3}$$

Erosion $(I \ominus S)$ on the other hand corresponds to the inverse operation, described as

$$I \ominus S = \{p | (p + q) \in I, \forall q \in S\}. \tag{3.4}$$

states that we only keep pixels $p \in I$ such that S_p fits inside I.

$$I \ominus S = \{p | S_p \subseteq I\}. \tag{3.5}$$

Both operation have the duality property, since erosion can be computed as a dilation of the background as

$$I \ominus S = \overline{\bar{I} \oplus S}. \tag{3.6}$$

and same duality can be applied to dilation,

$$I \oplus S = \overline{\bar{I} \ominus S}. \tag{3.7}$$

.

Results of both operations are exhibited in Figure 3.6.

Original Erosion Dilation

Figure 3.6: Comparison between Dilation and Operations

Both operation can be extended to operate in grayscale images, by the use of the max and min operations for a particular structuring element S,

$$\begin{aligned} I \oplus S_{(x,y)} &= \max_{u,v|\in S}\{I(x-u, y-v)\} \\ I \ominus S_{(x,y)} &= \min_{u,v|\in S}\{I(x-u, y-v)\}, \end{aligned} \tag{3.8}$$

and deriving for handling real values yields

$$\begin{aligned} I \oplus S_{(x,y)} &= \max_{u,v|\in S}\{I(x-u, y-v) + S_{(u,v)}\} \\ I \ominus S_{(x,y)} &= \min_{u,v|\in S}\{I(x-u, y-v) - S_{(u,v)}\}. \end{aligned} \tag{3.9}$$

Opening, Closing, Top Hat operations are derived from combinations of dilation's and erosion's operations in different sequence order. Combined with the proper structuring element, small image artifacts can be easily suppressed.

Opening operation corresponds to a **erosion** followed by a **dilation** (Equ 3.10), while **closing** corresponds to the opposite order of operations (Equ 3.11:

$$I \circ S = (I \ominus S) \oplus S \tag{3.10}$$

$$I \bullet S = (I \oplus S) \ominus S \tag{3.11}$$

Foreground structures that are smaller than structure element S can be removed by **opening** operations, while with **closing** operations, holes in the foreground smaller than S are filled.

In addition, two **Top-hat** transformations can also be defined, First, the white top-hat transformation, corresponding to the difference between the input image I and its **opening** by some structuring element S, $T_w(I,S) = I - I \circ S$, and second transformation, the black top-hat, corresponds to the difference between the **closing** and the input image I, $T_b(I,S) = I \bullet S - I$.

3.4 Pectoral Muscle Segmentation

Image segmentation is the division of an image into regions or categories that correspond to different objects or parts of then. Every pixel in an image is allocated to one of these categories. A good segmentation typically assigns to the same category, pixels that have

similar grayscale or multivariate values, creating a connected region. Good pectoral muscle segmentation enables to increase the robustness of the subsequent methods by removing non-relevant regions. To assess the potentiality and limitations of several segmentation methods based on region growing, intensity, graph and deep leaning are employed in pectoral muscle segmentation.

3.4.1 Background Removal

OTSU (Otsu, 1979) method finds a threshold that minimizes the weighted within-class pixel variance. It assumes that the histogram of the image is bi-modal and no spatial coherence or any notion of object structure exists. Images that exhibit bi-modal histograms are easily separable.

3.4.2 Region Growing

Region growing segmentation method (Adams and Bischof, 1994) relies on the idea that a group of pixels or sub-regions can be assigned into larger regions based on pre-defined criteria. The pixel aggregation starts with a defined seed point from were the corresponding regions will grow, appending to each seed those neighboring pixels that share similar properties such as gray level, texture or color. The process stops when no more pixels can be added. Region-based segmentation methods follow these basic premises:

- Completeness: $\bigcup_{i=1}^{n} R_i = R$ $\rightarrow$ The region must be complete, i.e, every pixel must be in a region.
- Connectedness: R_i is a connected region $i = 1, 2, ..., n$ $\rightarrow$ The points of a region must be connected in some sense.
- Disjointness: $R_i \bigcap R_j = \varnothing$ for all $i = 1, 2, ..., n$ $\rightarrow$ Regions must be disjoint.
- Satisfiability: $P(R_i) = TRUE$ for $i = 1, 2, ..., n$ $\rightarrow$ Pixels from a area must satisfy one common property P at least, i.e, any region must satisfy a homogeneity predicate P.
- Segmentability: $P(R_i \bigcup R_j) = FALSE$ for any adjacent region R_i and R_j $\rightarrow$ Different regions satisfy different properties, i.e, any two adjacent regions cannot be merged into single region.

In the specific case of mammogram images (Figure 3.4), the pectoral muscle border exhibits a very faded appearance regarding its neighborhood, causing difficulties to the Region Growing algorithm. To facilitate the task, histogram equalization is performed in the original

image to enhance pixels regions and the output of the region growing algorithm is subject to a dilation operation to fill gaps and smooth the final pectoral muscle contour (Figure 3.7).

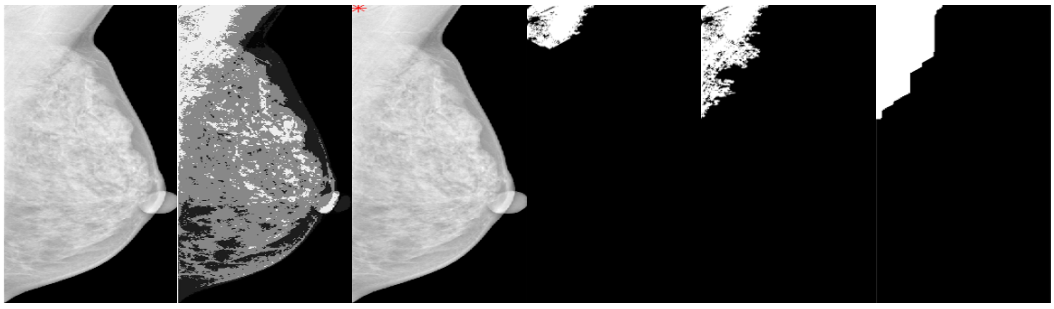


Figure 3.7: Region growing evolution inside pectoral muscle region.

3.4.3 Active Contours

Active Contours Models (ACM) or snakes Kass et al. (1988) corresponds to the minimization of an energy spline, guided by external constraint forces that pulls the spline towards region edges boundaries. Snakes are a generalist technique for matching a deformable model to an image region by means of energy minimization. The external forces are responsible for putting the snake near the desired local minimum. By representing the position of the snake parametrically $v(s) = (x(s), y(s))$, the energy can be written as

$$E^*_{snake} = \int_0^1 E_{snake}(V(s))ds = \int_0^1 E_{int}(v(s)) + E_{ext}(v(s)) + E_{con}(v(s))ds \tag{3.12}$$

where E_{init} represents the spline internal energy due to bending, E_{ext} the external acting forces and E_{con} the external constraint forces. The internal spline energy at a particular contour point $v(S)$ is evaluated as

$$E_{int} = \alpha(s)\left|\frac{\partial V}{\partial S}\right|^2 + \beta(s)\left|\frac{\partial^2 V}{\partial^2 S}\right|^2 \tag{3.13}$$

where $\alpha(s)$ controls to the first order term (elasticity), making the snake act like a membrane while the second order term is controlled by $\beta(s)$ assessing the stiffness, making the snake to act as a thin plate. The external energy describes how well the curves match the local image point. Considering a image $I(x, y)$, the gradient $\nabla I = (i_x, I_y)$ at any given point and the edge strength at pixel $(x, y) = |\nabla I(x, y)|$, the external energy of a contour point $v = (x, y)$ is given as

$$E_{ext}(v) = -\left|\nabla I(v)\right|^2 = -\left|\nabla I(x,y)\right|^2 \tag{3.14}$$

Now the total energy of a basic elastic snake becomes

$$E = \alpha \int_0^1 \left|\frac{\partial V}{\partial S}\right|^2 ds - \int_0^1 \left|\nabla I(v(s))\right|^2 ds. \tag{3.15}$$

To avoid the initial snake contour to be nudged in areas where it goes wrong, an extra external energy constraint term, E_{cont} is added, to pull nearby points towards or push them away (Equ 3.16) respectively

$$E_{pull} = -\int_0^1 \frac{r^2}{|V(s)-p|^2} \ , \ E_{push} = \int_0^1 \frac{r^2}{|V(s)-p|^2} \tag{3.16}$$

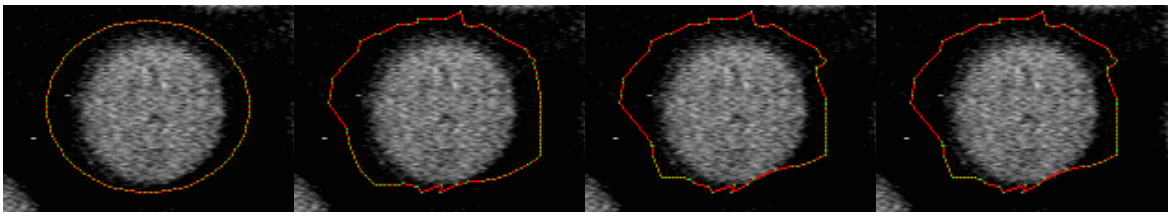
Now the final problem corresponds to the minimization of the total snake energy. Gradient descent can be employed to this task, by modeling a simple elastic snake energy as

$$E(x_0,...,x_{n-1},y_0,...,y_{n-1}) - \sum_{i=0}^{n-1} |I_x(x_i,y_i)|^2 + |I_y(x_i,y_i)|^2 + \alpha \sum_{i=0}^{n-1} (x_{i+1}-x_i)^2 + (y_{i+1}-y_i)^2. \tag{3.17}$$

The update equation for the whole snake is then defined as:

$$C' = C - \nabla E * \Delta t \tag{3.18}$$

where C correspond to the $E(x_0,...,x_{n-1},y_0,...,y_{n-1})$ components. The equation 3.17 calculates the energy gradient between the current boundary location and neighbors pixels, by performing expansions and contractions based on the gradient. Dynamic programming techniques for snake energy minimization can be also used (Mortensen and Barrett, 1998; Amini et al., 1990). Several interactions of the snake algorithm are exhibited on Figure 3.8.



Iteraction 1 Iteraction 10 Iteraction 50 Iteraction 100

Figure 3.8: Iteration of the Snake. (Image from [5])

[5] https://www.markschulze.net/snakes/

The main problem of snakes relies on the fact that is dependable on a set of numbers and spacing control points, making quite sensitive to the initialization point and noisy images. Improper choices may lead to situations were topological changes of objects not being followed.

3.4.3.1 Chan-Vese Model

The edge-based active contours are very sensitive to noise and usually fail to find the object boundaries in noisy images. Researchers started thinking on how to modify the stopping criteria designing a function that doesn't depend on the gradient, but some other property of the object in an image. Chan and Vese (2001) the approach uses the region properties to stop the curve at the object boundary. The main idea behind the model was to compute two energies (E_1 and E_2), such as

$$\begin{cases} E_1(V) = \int V_{inside} |I - C_1|^2 \delta_x \delta_y \\ E_2(V) = \int V_{outside} |I - C_2|^2 \delta_x \delta_y \end{cases} \tag{3.19}$$

where, V denotes the contour, I the image, C_1 and C_2 the average grayscale intensities inside and outside of the contour V, respectively, formulated as

$$C_1 = \frac{\int V_{inside} |I| \delta_x \delta_y}{|V_{inside}|} \ , \ C_2 = \frac{\int V_{outside} |I| \delta_x \delta_y}{|V_{outside}|} \tag{3.20}$$

As an illustrative example on how these energies work, assuming a grayscale image with an object with low intensity and the background with a high intensity(Figure 3.9). Based on the energies, four conditions can be derived:

- Contour is outside the object boundary, $E_1(v) > 0$ and $E_2(v) \approx 0$.
- Contour is inside the object boundary, $E_1(v) \approx 0$ and $E_2(v) > 0$.
- Contour is across the object boundary, $E_1(v) > 0$ and $E_2(v) > 0$.
- Contour is located on the object boundary, $E_1(v) \approx 0$ and $E_2(v) \approx 0$, (required condition).

As the fourth condition being the required condition, the model $E_1 + E_2$ becomes a region based energy minimization problem. To regulate the motion of the curve, Chan and Vese (2001) added the length term and the area term to the model. The final total energy is defined as:

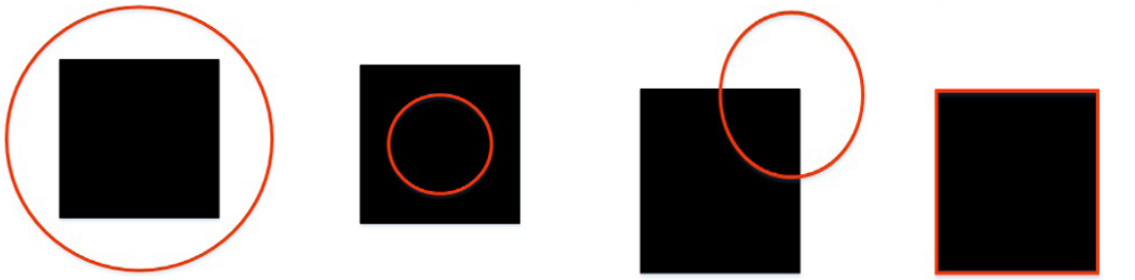


Figure 3.9: All possible curve conditions.

$$E_{CV} = \mu L(V) Q A_{inside}(v) + \lambda_1 E_1(v) + \lambda_2 E_2(V) \tag{3.21}$$

where $\mu, Q, \lambda_1, \lambda_2$ are the parameters that control the importance of each energy component. The length term smooths the contour by minimizing its length. The area term is used to accelerate the contour and helping in conditions when the initial contour is far from the object boundary.

3.4.4 Multi-Intensity Segmentation

Multi-intensity methods are simple intensity-based segmentation methods. The process starts by first reducing the number of gray-scale levels of the image and merging the higher levels into a single one forming a binary image. The obtained regions contain non-connected segments that are subject to region labeling. Considering the fact that labeling mechanism scans the image left to right, top to bottom is possible to attain only the first label numbers, corresponding to regions located at the top left of the image, (the location that encompasses the pectoral muscle region). Each of the selected levels is subject to morphological closing operations before being merged to fill open holes. The structuring element was set to 10 pixels with a circular shape. Figure 3.10 exemplified the performed steps sequence.

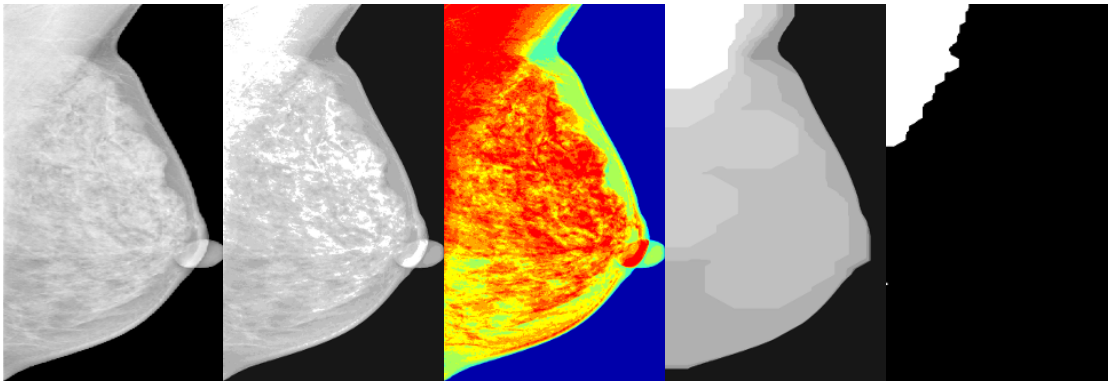


Figure 3.10: Multi Intensity segmentation stages.

3.4.5 Shortest Path Polar Coordinates (SPPC)

SP segmentation methods (Cardoso et al., 2010) explore the fact that a image can be represented as a grid, enabling to construct a graph with pixels acting as nodes and edges connecting the neighbour pixels. A graph $G = (C, A)$ is composed by a set of nodes V and a set of arcs (p, q), A with $p, q \in V$. A graph is weighted if a weight $w(p, q)$ is associated to each of the arcs. Weight of the arc $w(p, q)$ are set as function of pixels and their relative positions. A path from a vertex v_1 up to vertex v_n is a list of unique vertices $v_1, v_2, \cdots, v_n$ with v_{n-1} and v_i corresponding to neighbour pixels. The total cost of a path corresponds to the sum of all arcs weights among the path $\sum_{i=2}^{n} w(v_{i-1}, v_i)$. A path from a source vertex v to a target vertex u is said to be the `shortest path`, yielding the total minimum cost of all v-to-u paths. The distance between a source vertex v and a target vertex u on a graph, $d(v, u)$ corresponds to the total cost of the shortest path among those two vertex. A path from a source vertex v to a sub-graph is said to be the shortest path between v and Ω if its total cost is the minimum among all v-to-$u \in \Omega$ paths. The distance from a node v to a sub-graph $\Omega, d(v, \Omega$ corresponds to the total cost of the shortest path between v and Ω as

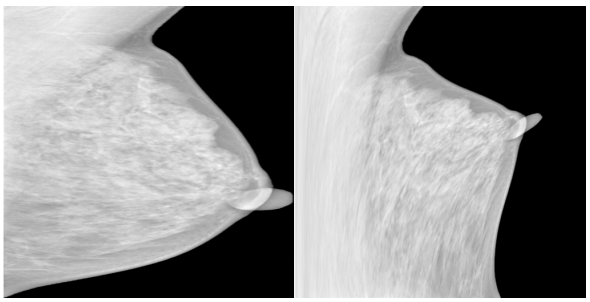
$$d(v, \Omega) = \min_{u \in \Omega} d(v, u) \tag{3.22}$$

A path from a sub-graph Ω_1 to sub-graph Ω_2 corresponds to the shortest path between Ω_1 and Ω_2 if the total cost is the minimum among all $v \in \Omega_1$-to-$u \in \Omega_2$ paths. The distance from sub-graph Ω_1 to sub-graph $\Omega_2, d(\Omega_1, \Omega_2)$ is the total cost of the shortest path between Ω_1 and Ω_2 with distance

$$d(\Omega_1, \Omega_2) = \min_{v \in \Omega_1, u \in \Omega_2} d(v, u) \tag{3.23}$$

In graph theory, the shortest-path problem seeks the shortest path connecting two nodes. Efficient algorithms are available to solve this problem, such as the well-known Dynamic Programming (DP) algorithms like Dijkstra algorithm Dreyfus (1969).

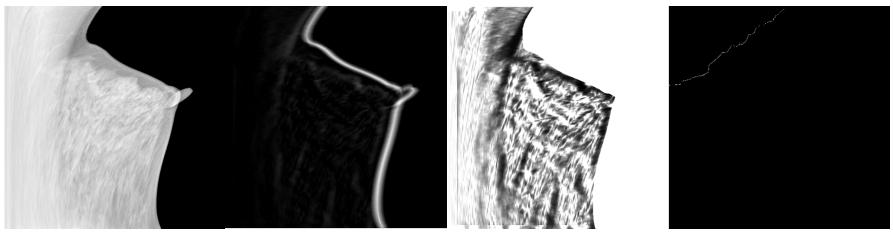
The main difficulty with searching for the SP between the top row and left column where pectoral muscle is located, relies on the fact that small paths, near the top-left corner, are naturally favored since. To circumvent this problem, the image is transformed into polar coordinates, Figure 3.11.



Original Polar

Figure 3.11: Original image on the left and polar transformed image on the right.

The center of coordinates is assumed to be the top-left image corner. On this new coordinate system, the path to search becomes now a minimum path search between the top and bottom rows. After polar transformation, a horizontal and vertical Prewitt kernel (Prewitt, 1970) can be employed to emphasize pectoral muscle edges. The resulting gradient image can now be considered as a weighted graph with pixels acting as nodes and edges connecting neighboring pixels. Figure 3.12 summarizes the main steps.



Polar Edge Cost Original Line

Figure 3.12: Pectoral muscle segmentation stages.

3.4.6 Encoder-Decoder Architecture (U-net)

U-net initially proposed by Ronneberger et al. (2015), is a network architecture for fast and precise segmentation of images. The main idea consists in extending a usual contracting CNN, by replacing pooling operators with up-sampling operators, increasing the resolution

of the output. For region localization, high-resolution features from the contracting path are combined with the up-sampled output, leading to a more precise output avoiding the checkerboard problems. This new successive convolution layer is able to learn how to assemble a more precise output based on this information.

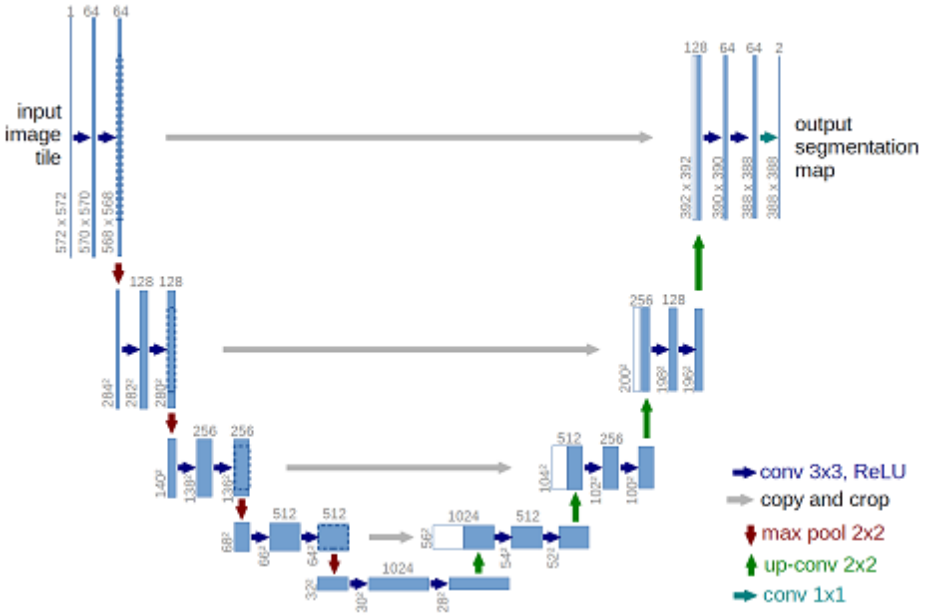


Figure 3.13: U-net architecture (example for 32×32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The $x - y$-size is provided at the lower left edge of the box. White boxes represent copied feature maps. (Image from Ronneberger et al. (2015)).

A particular modification in the architecture relies on the up-sampling path, where exist a large number of feature channels, allowing the network to propagate the context information into higher resolution layers. As consequence, the expansive path (right side) and the contracting path (left side) are approximately symmetric, forming a u-shaped architecture (Figure 3.13). Contrary to the CNN, the network does not contain fully connected layers, attaining only the valid parts of each convolution layer, enabling to create a segmentation map that contains only pixels from where its full context is available from the input image.

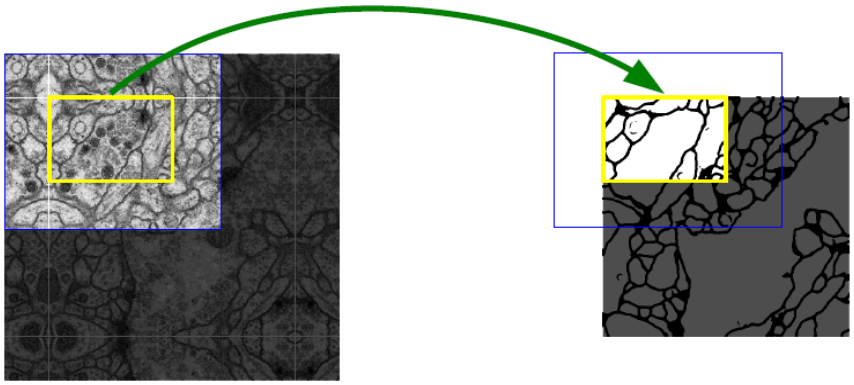
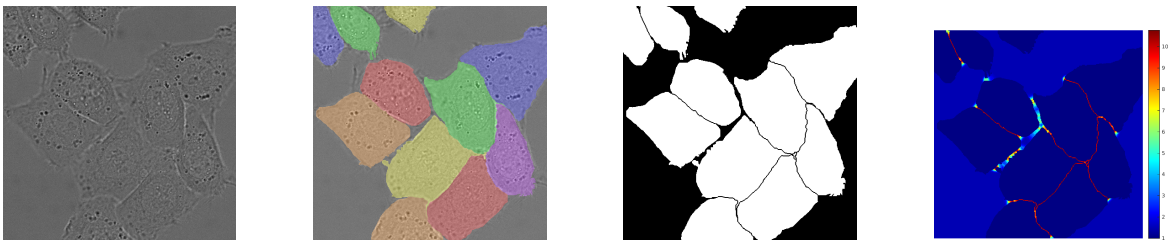


Figure 3.14: Overlap-tile strategy for seamless segmentation of arbitrary large images. Prediction of the segmentation in the yellow area, requires image data within the blue area as input. Missing input data is extrapolated by mirroring. (Image from Ronneberger et al. (2015)).



(a) Original (b) Overlay with GT (c) Generated mask (d) Pixel-wise loss

Figure 3.15: HeLa cells on glass semantic segmentation. (Image from Ronneberger et al. (2015)).

The use of overlap-tile strategy allows seamless segmentation of arbitrarily large images (Figure 3.14). The contracting path consists in repeated application of two 3×3 convolutions (un-padded convolutions) followed by Rectified Linear Unit (ReLU) and a 2×2 max -pooling operation with stride of 2 for down-sampling. On each down-sampling step, the number of feature channels is doubled. The inverse path is taken in every step by the expansive path. This path consists in an up-sampling of the feature map followed by a 2×2 up-convolution, halving the number of feature channels, a concatenation with the corresponding cropped feature map from the contracting path and two 3×3 convolutions, each followed by a ReLU. Cropping stage is necessary due to the loss of border pixels in every convolution stage. The final layer consists in a 1×1 convolution, used to map each of the 64 components feature vectors into the desired number of classes, forming a final network with 23 convolutional layers. To allow a seamless tiling of the output segmentation map, the selection of the input tile size must be such that all 2×2 max-pooling operations are applied to a layer with an even x- and y-size.

Input images, corresponding region segmentation masks, and Stochastic Gradient Descent

(SGD) are used to train the network. The un-padded convolutions transform the output image into a smaller version of the original input with a constant border width. To maximize Graphical Processing Unit (GPU) memory is common to favor large input tiles over a large batch size, reducing the batch size towards a single image. The energy function is computed by a pixel-wise soft-max function over the final feature map, combined with the cross-entropy loss function. The soft-max energy function is defined as:

$$p_k(x) = \frac{\exp a_k(x))}{\sum_{k'=1}^{K} \exp(a_{k'}(x))} \tag{3.24}$$

where $a_k(x)$ denotes the activation on the feature channel k at the pixel position $x \in \Omega \subset \mathbb{Z}^2$. K corresponds to the number of classes with $p_k(x)$ being the approximated maximum-function, where $p_k(x) \approx 1$ for k, containing the maximum activation $a_k(x)$, while $p_k(x) \approx 0$ for all the other k. On each position the cross entropy penalizes the deviation of $p_{\ell(x)}(x)$ from 1 using

$$E = \sum_{x \in \Omega} w(x) \log(p_{\ell(x)}(x)) \tag{3.25}$$

where $\ell : \Omega \to \{1, \cdots, K\}$ corresponds to the true label for each pixel and $w : \Omega \to \mathbb{R}$ the introduced weight map to increase the importance of particular pixels during training. For each GT, a weight map is computed to compensate the different pixel frequency from certain classes on the training data set, forcing the network to learn the small separation borders (Figure 3.15). The separation border is obtained using morphological operations, and the the weight map is computed as

$$w(x) = w_c(x) + w_0 * \exp -\frac{(d_1(x) + d_2(x))^2}{2\sigma^2} \tag{3.26}$$

where $w_c : \Omega \to \mathbb{R}$ corresponds to the weight map that balances the class frequencies, $d_1 : \Omega \to \mathbb{R}$ denotes the distance to the border from the nearest cluster cell and $d_2 : \Omega \to \mathbb{R}$ the distance to the second cell. w_0 corresponds to the initial weight map value while σ expresses pixel deviation.

To circumvent the sparse training data, excessive data augmentation is commonly used by applying elastic deformations to the available training set, allowing the network to learn invariance to such deformations, resulting in a network that generalizes better.

3.4.7 Experiments and Results for Pectoral Muscle Segmentation

For pectoral muscle segmentation, all images are assumed to been subject to orientation homogenization and attain only breast region. The previously described methods, SP in polar coordinates, snakes, region growing, intensity (grey-level), and deep learning semantic segmentation using U-net with a focus in pectoral muscle region are implemented and evaluated against baseline and state of the art methods.

For the U-net described in Table 3.1, images were separated in training, validation and test sets using a split of 70%, 15% and 15% respectively, combined with data augmentation operations, such mirroring and angle rotation $0°, 90°, 180°$ and $270°$ degrees , increasing the training set by a factor of 4×2. All images in U-net were resized to 512×512 for training and evaluation purposes. ADAM was the selected optimizer with learning rate $\lambda = 2 \times 10^{-4}$ using binary cross-entropy. The number of epochs was set to 40. In order to improve model convergence, background reduction plus contrast normalization to highlight brighter areas (muscle and lesions) was performed by setting to zero any pixel below the 0.01 value (background). Intensity was normalized on each image individually before input, with pixels distribution linearly scaled to cover the entire intensity range $[0 - 1]$.

Table 3.1: Description of the U-Net architecture used for segmentation. All Convolutional are followed by a ReLU activation. The output layer has a *sigmoid* activation function for binary classification. Note: ReLU layers were omitted from description simplicity.

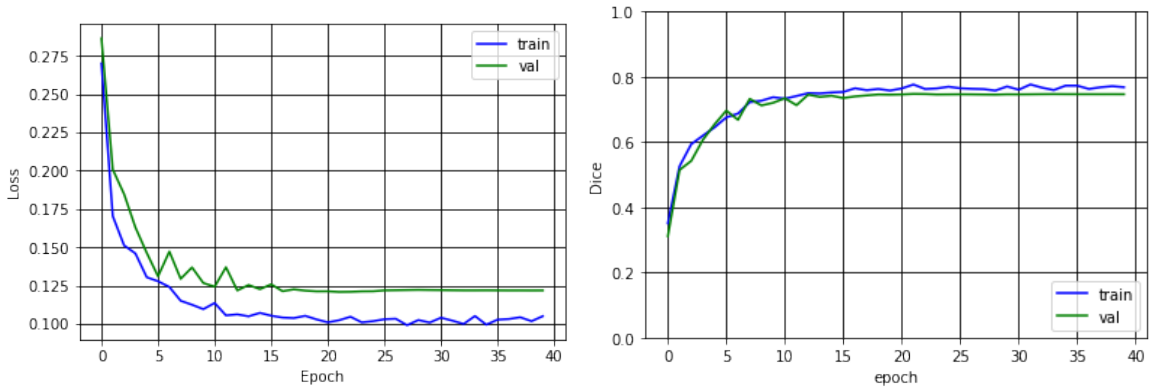
Table 3.2: Down part

Layer	# Filters	Filter Size
Input	512×512	-
Convolutional	64	3
Convolutional	64	3
MaxPolling	2×2	2
Convolutional	128	3
Convolutional	128	3
MaxPolling	2×2	2
Convolutional	256	3
Convolutional	256	3
MaxPolling	2×2	2
Convolutional	512	3
Convolutional	512	3
DropOut	0.5	-
MaxPolling	2×2	2
Convolutional	1024	3
Convolutional	1024	3
DropOut	0.5	-

Table 3.3: Up Part

Layer	# Filters	Filter Size
Convolutional	512	3
Convolutional	512	3
UpSampling	2×2	-
Convolutional	256	3
Convolutional	256	3
UpSampling	2×2	-
Convolutional	128	3
Convolutional	128	3
UpSampling	2×2	-
Convolutional	64	3
Convolutional	64	3
UpSampling	2×2	-
Convolutional	2	3
Convolutional	1	1
Sigmoid	-	-

Convergence and Dice Coefficient (DC) metrics of the U-net model are exhibited on Figure 3.16,



Loss - Train/Validation

Dice - Train/Validation

Figure 3.16: Loss and Dice coefficient during training.

The final model yielded a DC of 0.733 for validation set.

For SP in polar coordinates, an exponential law for weight creation was defined to enhance pectoral muscle contour from other regions, with weight w to a arc being determined by the gradient values of the two incident pixels and 4-neighbour pixels, p and q expressed as

$$\hat{f}(g) = f_l + (f_h - f_l)\frac{\exp\left((255 - g)\cdot\beta\right) - 1}{\exp\left(255\cdot\beta\right) - 1} \tag{3.27}$$

with $f_h, f_l, \beta \in \mathbb{R}$ set to constants values $f_h = 32, f_l = 3, \beta = 0.0208$ and g the minimum of the gradient on the two incident pixels. Additional directional cost parameters $C_{right} = 3.3$ and $C_{left} = 1.1$ to modulate the graph cost towards the left side of the image. All images, independently of the original size, were resized to 1024×1024 for polar cost computation and converted back to original size without degrading final result. Additional processing was made to avoid the influence of the external factors such as outside breast contour. Situations were the outside breast border that present strong edge response close to the muscle region, misleading the determination of the SP calculation, and to minimize this influence a eroding operation on the outside breast contour was made.

For region growing, the structural element was set to 10 pixel radius with a tolerance of $tol = 0.1$, enabling to attain the rough boundary and reconstruct pectoral muscle contour. For active contours segmentation, the λ and the number of iterations were set to 2 and 1000 respectively, in order to have sufficient iterations in order to attain larger pectoral muscle regions.

For intensity segmentation, the number of gray levels was set to 9 with the threshold level set to 7. This enabled to attain brighter regions that also contain pectoral muscle region.

before merging upper level into a single one, each of the upper levels is subject to close operations with a circular structure element of size 10. After merging region labeling using 4-neighbor scheme is performed attaining only the first label that encompasses the pectoral muscle area.

A pairwise contour comparison of all methods is presented in Figure 3.17. All methods were compared with the GT gathered from the INbreast databases with contour metrics normalized regarding the Region of Interest (ROI) diagonal.

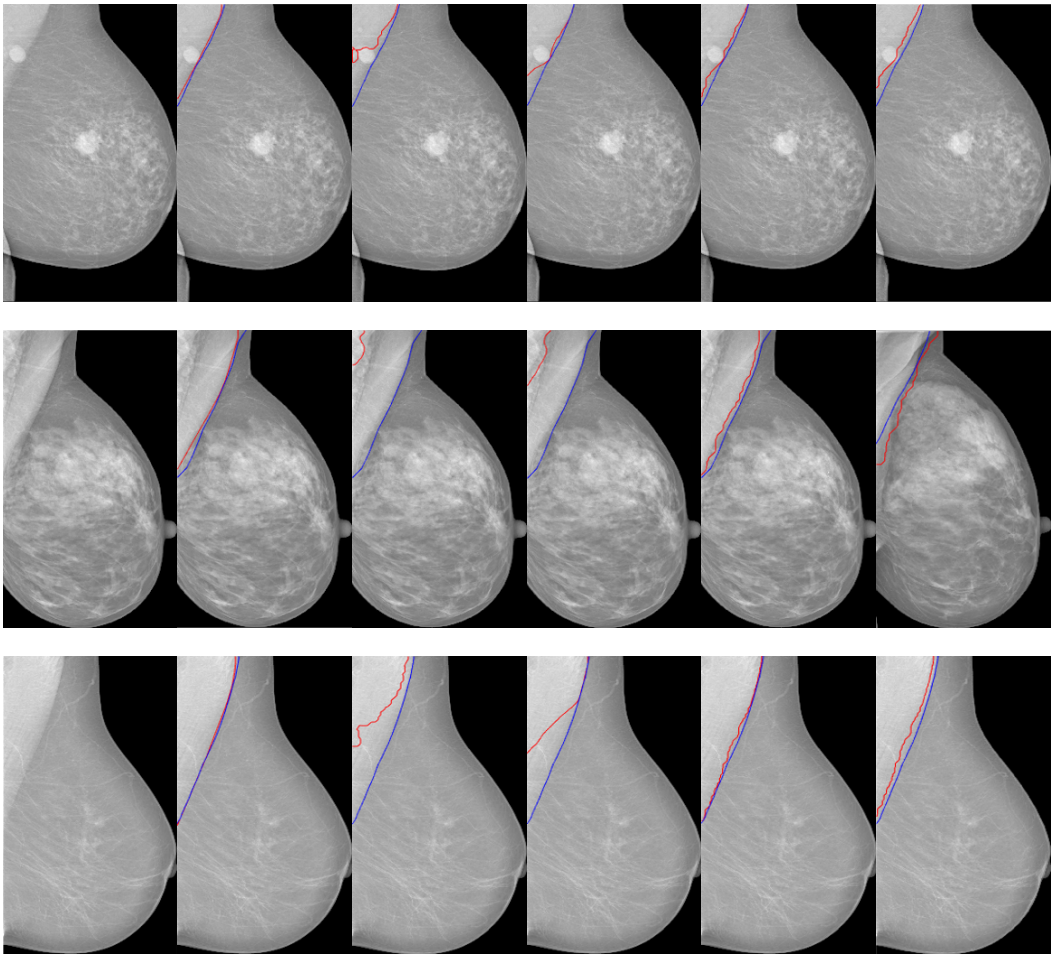


Figure 3.17: Example of the implemented segmentation methods (Blue - GT and Red - Detection).

Table 3.4 presents the performance for all methods. Besides the proposed methodologies, a baseline segmentation method is introduced, corresponding to the straight line that connects the mean of GT start position to the mean end position of all images containing pectoral muscle. Setting a state of the art base comparison for pectoral muscle segmentation, Taghanaki et al. (2017) combines geometric rules with a region growing algorithm to support the

segmentation of all types of pectoral muscles (normal, convex, concave, and combinatorial), yielding a DICE similarity coefficient of 0.972(0.003) in INbreast database.

Table 3.4: Overall results in the position of the muscle boundary. Results are in mean (std).

Method	AD	AMED	HD	AOM	CM	DICE
SotA	-	-	-	-	-	0.972(0.003)
Baseline	0.049(0.004)	0.059(0.006)	0.121(0.011)	0.577(0.034)	0.712(0.025)	0.727(0.096)
SPPC	0.062(0.021)	0.065(0.015)	0.161(0.029)	0.735(0.036)	0.822(0.021)	0.799(0.028)
RG	0.204(0.017)	0.213(0.017)	0.402(0.031)	0.460(0.055)	0.535(0.040)	0.543(0.066)
Levels	0.272(0.086)	0.378(0.104)	0.743(0.116)	0.287(0.029)	0.436(0.025)	0.340(0.069)
Active	0.265(0.074)	0.366(0.170)	0.748(0.193)	0.226(0.070)	0.581(0.072)	0.298(0.103)
U-net	0.187(0.056)	0.231(0.060)	0.422(0.076)	0.704(0.085)	0.722(0.064)	0.723(0.061)

AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better), while AD, AMED and HD are measures of pixel error (the lower the better).

Intensity (grey-level) and active contours presented non-satisfiable results due to the fact they act in the intensity domain, difficulting the determination of the best parameters. Comparing the baseline with all other methods, SP proved to be the most effective, however pectoral regions containing diffuse tissues lead to the SP algorithm to pick wrong edges instead of real muscle contour increasing the contour error metrics (large STD). U-net proved to be robust regarding diffuse tissues, but the over-segmentation of pectoral muscle region degraded the performance. A combination of U-net followed by a refinement stage using SP in Polar Coordinated can increase the overall performance of the pectoral muscle segmentation stage, avoiding that diffuse tissues being picked by the SP in Polar Coordinates, enabling that the correct muscle contour endpoints being properly selected among with the corresponding path.

3.5 Mass Lesion Detection

Having screened out normal mammograms, the following task typically involves looking for suspicious regions in the mammogram. Two types of findings can be present in mammogram images, calcifications and masses. Due to their differences, specialized detection systems are typically developed for each of findings. For the task of automatically detect masses, three computer vision methods (Saliency Maps, Watershed and Iris Filter) are described below, followed by a FP reduction stage consisting in a Support Vector Machine (SVM) classifier fitted with contour, texture and statistical describing features.

3.5.1 Salience Maps

Saliency Map (Achanta et al., 2008) of an image corresponds to pixel's that exhibit unique characteristics when compared with remainder pixels. This enables to represent an image

into something more meaningful and easier to analyze. For example, if a pixel exhibits a high grey level quality on an image, that pixel's quality will present a saliency map in an obvious way. Spatial attention models are commonly employed to extract visual signals, such as intensity, color or texture, forming the saliency map. A saliency value of a pixel I_k on image I is defined as:

$$SalS(I_k) = \sum_{\forall I_i \in I} ||I_k - I_i||, \tag{3.28}$$

where I_i is the color value of each pixel on image I, with range between $[0 - 255]$, while $||.||$ represents the color distance metric. Expanding Equation 3.28 results

$$SalS(I_k) = ||I_k - I_1|| + ||I_k - I_2|| + \cdots + ||I_k - I_N||, \tag{3.29}$$

with N corresponding to the total number of pixels on image I. Let $I_k = a_m$ and Equation 3.29, the terms I_i can be rearranged as

$$\begin{aligned} SalS(I_k) &= ||a_m - a_0|| + \cdots + ||a_m - a_1|| + \cdots + \cdots, \\ SalS(a_m) &= \sum_{n=0}^{255} f_n ||a_m - A_n||, \end{aligned} \tag{3.30}$$

where f_n corresponds to pixel frequency value (a_n) on image I. This frequency can be expressed in the form of histograms. Since $a_n \in [0, 255]$, the color distance metric $||a_m - A_n||$ is bounded also to the $[0 - 255]$ range. With this fixed range, a distance map D can be obtained prior to the saliency map computation. In this map, the element $D(x, y) = ||a_x - a_y||$ corresponds to the color difference between a_x and a_y, (Figure 3.18).

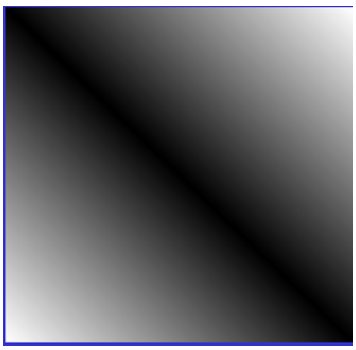
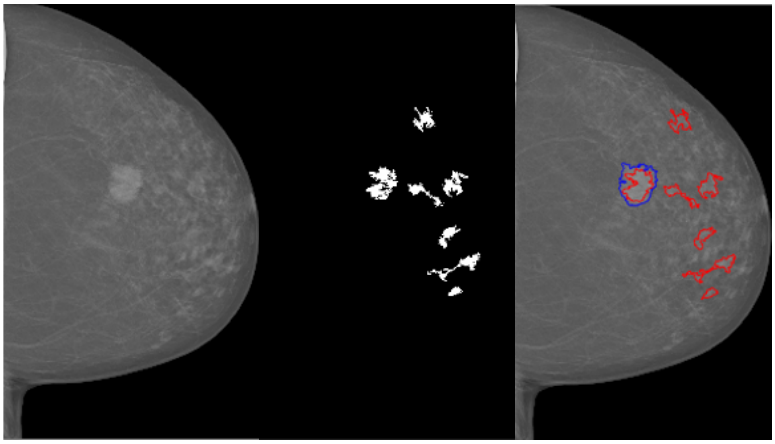


Figure 3.18: The distance map between the gray-level color values. Brighter elements represent larger distance values (Image taken from Zhai and Shah (2006).

Given a histogram $f(\cdot)$ and the corresponding color distance map, $D(\cdot, \cdot)$, the saliency value for a pixel I_k is given as

$$SalS(I_k) = SalS(a_m) = \sum_{n=0}^{255} f_n D(m,n). \tag{3.31}$$

Alternatively, computation of the saliency values of all the image pixels using Equation 3.28 is possible, requiring only the saliency values of colors $a_i, i = 0, \cdots, 255$ to generate the final saliency map. Example of the pixel-level spatial saliency computation is shown in Figure 3.19.



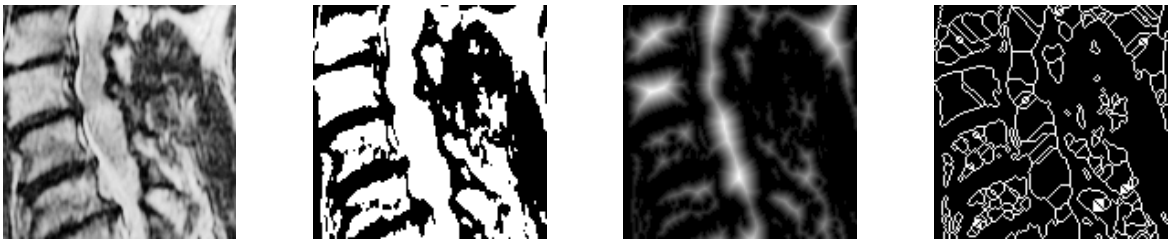
Original Saliency Map Contour

Figure 3.19: An example of the spatial saliency computation (Blue - GT, Red - Detections).

3.5.2 Watershed

In geography, a watershed is a ridge that divides areas that are drained by different river systems. A catchment basin in this sense resembles an area from which rainfall flows into a reservoir. Watershed segmentation (Beucher and Meyer, 1992) applies these key ideas to gray-scale images, enabling to solve a variety of image segmentation problems. Considering a gray-scale image as a topological surface where the values of $f(x, y)$ are interpreted as heights, the watershed is able to find the reservoirs and ridgelines contained in a grayscale image. The concept starts by transforming the input image into another, were reservoirs are the objects or regions to be identified. Two main transforms can be employed, (1) First and the more common, the use of a Distance Transform representing the distance from each pixel with a value of 1 to the nearest non-zero pixel value. By thresholding, a gray-scale image using OTSU(3.4.1) and taking its complement, creates a binary image that highlights areas to be captured. Applying the watershed distance transformation to the image and assign all zero pixels of the complementary image to $-\infty$, results in a labeled matrix that identifies watershed regions with integer elements ≥ 0. Zero values identify image contours while non-zero elements belong to the watershed regions. A final complementing operation is performed

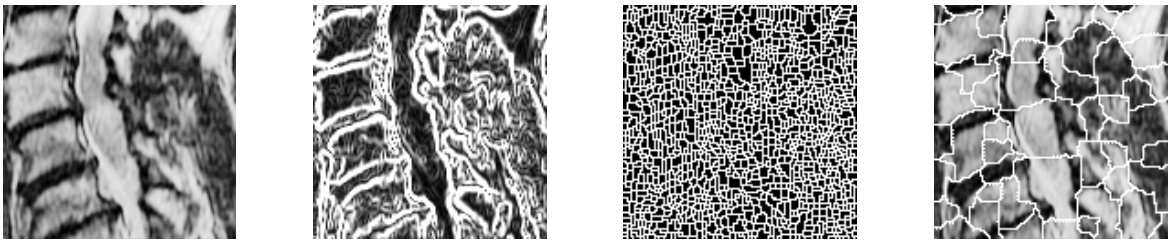
by assigning all 1 values to zero elements and 0 values to all non-zero elements (Figure 3.20).



Original | Gradient magnitude | Watershead transform | Smooth watershead

Figure 3.20: Watershead distance transform segmentation (Image from Gavlasová et al. (2006)).

(2) Second, a gradient-based watershed segmentation method is also possible. The process starts by obtaining the magnitude gradient of the original image using linear filtering methods, such as Sobel, and compute the watershed transform of the corresponding gradient. To avoid over-segmentation, the gradient image must be smoothed before the watershed transformation. Typical morphology operations such as closing and opening are commonly used to refine the final segmented region. Figure 3.21 shows the method applied to a medical image.



Original | Binary | Distance transform | Watershead transform

Figure 3.21: Watershead gradient transform segmentation (Image from (Gavlasová et al., 2006)).

3.5.3 Iris Filter

Coin Filter (CF) and Iris Filter (IF) proposed by Kobatake and Hashimoto (1999) and analyzed in detail by Esteves et al. (2012), belong to the category of Local Convergence Filters (LCF) and have been shown to be quite robust to identify and segment regions that present low contrast. LCF evaluate the degree of convergence of the gradient vectors within a local area (support region) toward a pixel of interest (area central location). This degree of convergence is related to the distribution of the directions of the gradient vectors and not to their magnitudes. For evaluating the convergence of each coordinate (x, y) in an image

(discarding borders), a 2D discrete image space gradient orientation is assumed regarding a convergence region support filter, defined as:

$$\alpha(x, y, \theta_i, m) = \tan^{-1}\left(\frac{\delta I(x_o, y_o)\delta x}{\delta I(x_o, y_o)\delta y}\right) \tag{3.32}$$

with $x_o = x + m * \sin(\theta i)$ and $y_0 = y + m * \cos(theta_i)$ where I corresponds to image, (θ_i, m) polar coordinates within support region, $\frac{\delta I}{\delta x}$ and $\frac{\delta I}{\delta y}$ row and column wise derivative respectively. Support region polar coordinates are defined by m, being measured in pixels and $theta_i$, as result of radial sampling, defined as:

$$\theta_i = \frac{2\pi}{N}(i - 1) \tag{3.33}$$

where N is the number of radial directions to be evaluated. The convergence coordinates (θ_i, m) is defined using the cosine between the polar direction θ_i and the image gradient for coordinate (x, y, θ_i, m) as

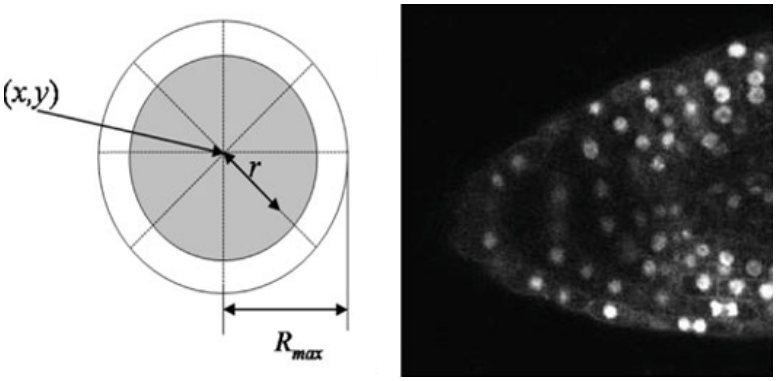
$$CI(x, y, i, m) = \cos(\theta_i - \alpha(x, y, \theta_i, m)) \tag{3.34}$$

with the overall convergence obtained by summing all the individual convergence from Equation 3.34.

Convergence index filter or COIN filter (CF) assumes a circle with variable radius (Figure 3.22) as support region to search of the maximum convergence value inside a limited radius R_max. The CF filter responses within a circle with varying radius are given as

$$CF(x, y) = \max_{0 \le r \le R_{max}} \frac{1}{N * r} \sum_{i=0}^{N-1} \sum_{m=1}^{r} CI(x, y, i, m) \tag{3.35}$$

where r is the radius of the circle of the support region that varies from 0 to R_max, N is the number of radial directions where the convergence is evaluated and $CI(x, y, i, m)$ defined in Equation 3.34. The result of applying Equation 3.35 on image (Fig 3.22 , results in a image (Figure 3.22.



(a) COIN filter (b) Original (c) CF responses

Figure 3.22: Schematic of the filter support region of the COIN filter (Support region as grey), Original image and CF responses.

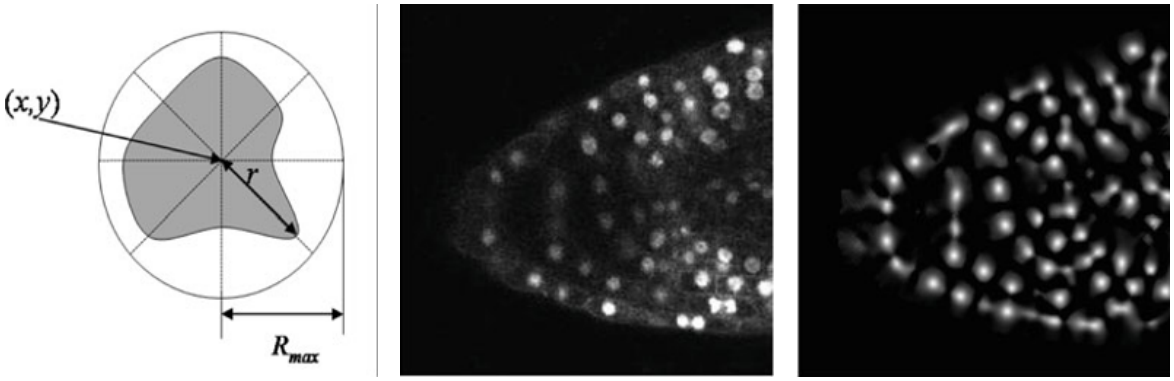
The maxima of such response indicate locations of interest. For each of the filters maxima, the radius of the corresponding support region can be obtained as

$$R_{shape}(x,y) = argmax_{0\leq r\leq R_{max}} \left[\frac{1}{N*r} \sum_{i=0}^{N-1} \sum_{m=1}^{r} CI(x,y,i,m) \right] \tag{3.36}$$

with R_shape being the radius of the support region (x,y) with the highest convergence.

IRIS filter (IF) is an evolution of the CF filter to handle a more diverse range of local convergence areas. The IF filter adapts the scan radius of its support region for each of the N directions, maximizing convergence for each radial direction independently, enabling to detect non-circular shapes (Figure 3.23). The convergence evaluation then becomes

$$IF(x,y) = \frac{1}{N} \sum_{i=0}^{N-1} \left[\max_{0\leq r\leq R_{max}} \frac{1}{r} \sum_{m=1}^{r} CI(x,y,i,m) \right] \tag{3.37}$$



(a) IF filter (b) Original (c) IF responses

Figure 3.23: Schematic of the filter support region of the IF filter (Support region as grey), Original Image and IF responses.

3.5.4 FP Reduction

Detection methods while able to extract ROI regions, also present a high number of FP's. To refine the detection stage, a FP reduction stage is can be added to reduce false detection's. In the particular case of mass detection, the use of a SVM classifier trained with contour and pixels features assigned to true and false regions labels can be used to reduce the FP rate and remove undesired detection's. The simplicity of the SVM classifier makes suitable for this task since the choice of the kernel and miss-classification cost enable to handle a reduced number of features and data. The SVM can be trained using several contour features extracted from the current detected areas, labeled as true if the area of the detection overlaps the GT of the mass contour of false if not.

3.5.5 Experiments and Results for Detection of Suspicious Mass Lesions

For the mass detection thee methods were implemented, they consist in Graph-Based Visual Saliency, Watershed and Iris Filter with $R_{max} = 30$ and $R_{max} = 50$ and later combined with a stage for FP reduction. Figure 3.24 shows earlier detections of the three methods.

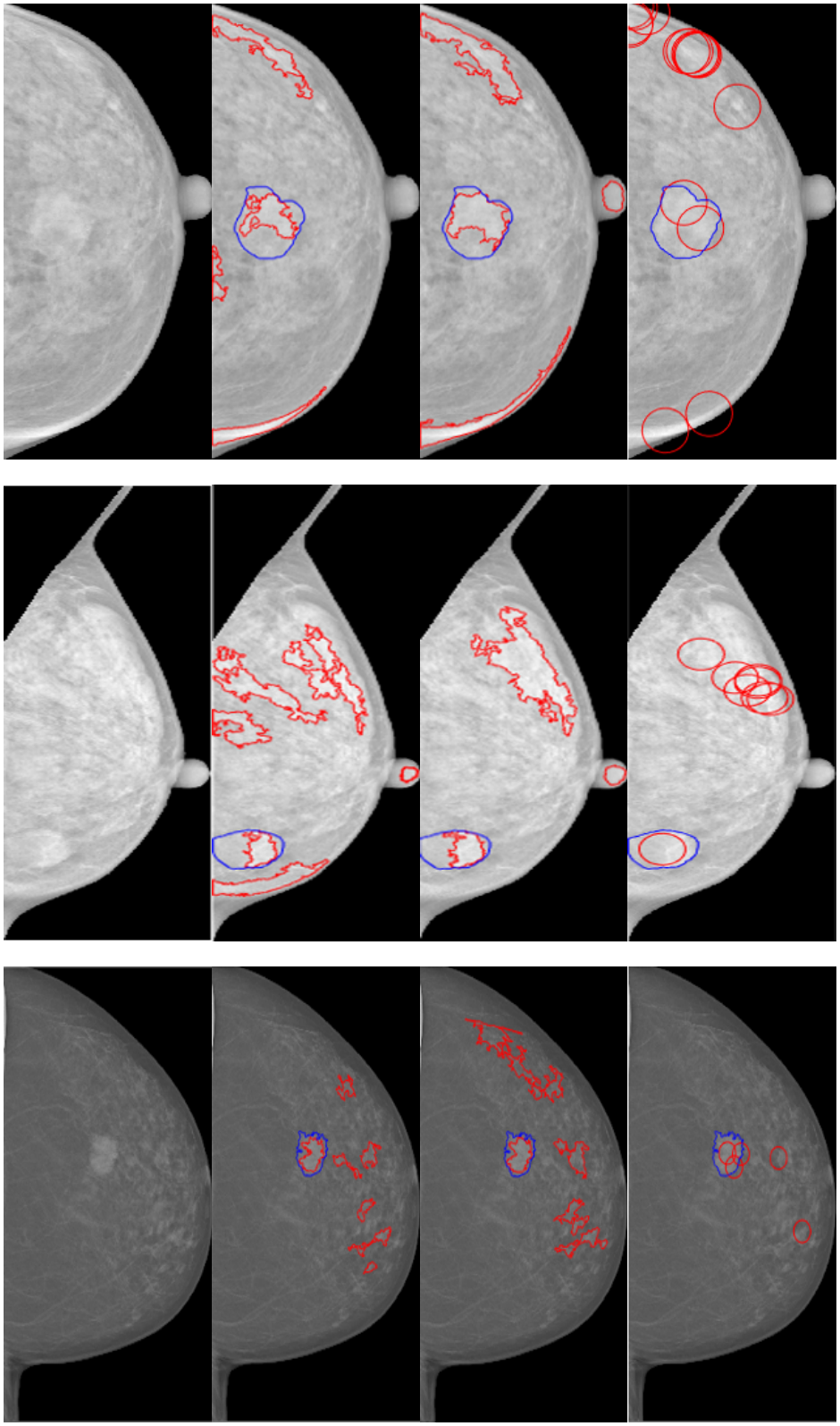


Figure 3.24: Example of the detection's of suspicious mass lesions using saliency, watershead and iris methods (Blue - GT and Red - Detection).

The Table 3.5 presents the computed metrics for the earlier detection's. For state of the art comparison, a recent work proposed by Dhungel et al. (2015), based on a cascade of CNN was selected, yielding an final $TP_r = 0.96(0.03)$. Region accuracy metrics were considered only if the detection overlap the GT, otherwise was set to zero.

Table 3.5: Performance evaluation of detection of suspicious mass lesions. Results mean (std).

Method	FP	TP_r	AOM	CM	DICE
SotA.	-	0.960(0.030)	-	-	-
Saliency	5(0.124)	0.673(0.064)	0.396(0.075)	0.560(0.091)	0.520(0.102)
Watershead	6(0.198)	0.663(0.095)	0.341(0.081)	0.521(0.101)	0.460(0.127)
Iris $R = 30$	13(1.265)	0.686(0.053)	0.245(0.092)	0.360(0.124)	0.432(0.139)
Iris $R = 50$	11(0.994)	0.673(0.064)	0.321(0.093)	0.414(0.119)	0.476(0.125)

FP(False Positives - lower the better), TP_r=Sens= $\frac{\#TP}{\#TP+\#FN}$ (Detection Rate/Sensibility - higher the better). AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better).

The Saliency presents a lower FP rate than Watershed method. Iris Filter presented a higher TP_r but with a higher FP. In terms of regions region metrics increasing the radius from $R_{max} = 30$ to $R_{max} = 50$ reduced the number of FP while increasing region metrics, resulting in more mass lesion area being contained on the detection.

To reduce the FP rate and remove undesired detection's, a SVM classifier was trained using several contours and texture features extracted from the current detection's patches images, labeled as true if the area of the detection overlaps the GT of the mass contour. The Table 3.6 summarizes the extracted features to characterize mass detected regions.

Table 3.6: Summary of the shape features that were selected for FP rejection.

Feature	Short Acronym
Eccentricity (Vadivel and Surendiran, 2013)	ECT
Extent	Ext
Dispersion (Vadivel and Surendiran, 2013)	Dp
Circularity (Vadivel and Surendiran, 2013)	Circ
Major Axis Length	MJl
Extent	Ext
Energy	Ener
Min value	Min
Max value	Max
Mean	Mean
Median	Median
Standard Deviation	Std
Grey-level difference matrix (Moura and López, 2013)	GLDM
Grey-level run length (Moura and López, 2013)	GLRL

The SVM was fitted with an RBF Kernel and trained on randomly selected subset con-

taining 75% of the whole dataset using 10 k-folds cross-validation with parameters C and α ranging from $[0, 10]$ and $[0.1, 2]$ respectively. The optimal final parameters were $C = 2$ and $\alpha = 0.2$. The choice of SVM was due to its simplicity and good performance when using a small range of features. Results after FP rejection are summarized in Table 3.7 and examples presented in Figure 3.25.

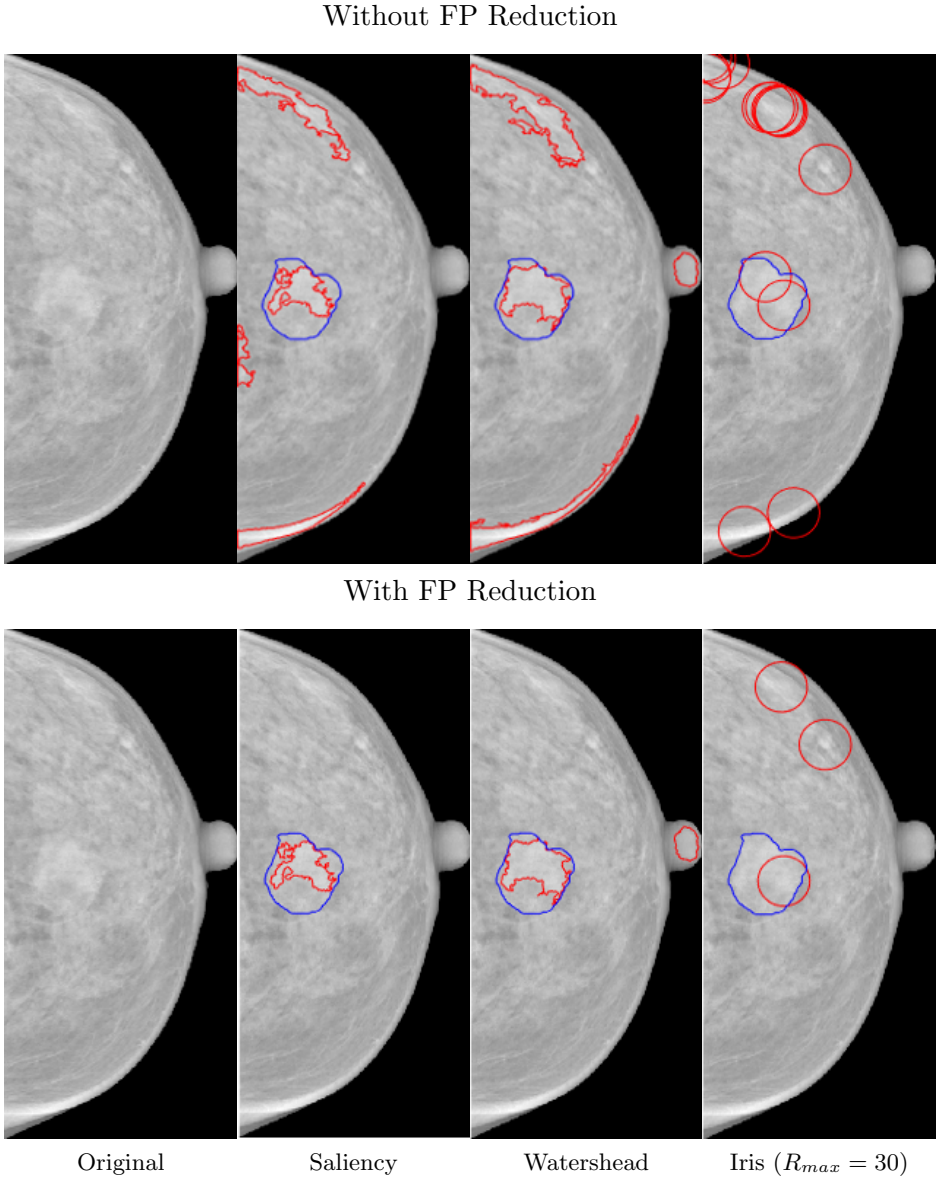


Figure 3.25: Example of the FP detection's reduction. (Blue - GT and Red - Detection).

Table 3.7: Performance evaluation of detection of suspicious mass lesions with FP rejection with SVM classifier. Results mean (std).

Method	FP	TP_r	AOM	CM	DICE
SotA.	-	0.960(0.030)	-	-	-
Saliency	2(0.094)	0.645(0.084)	0.372(0.069)	0.549(0.089)	0.524(0.097)
Watershead	3(0.031)	0.635(0.104)	0.331(0.082)	0.521(0.098)	0.443(0.123)
Iris $R = 30$	9(0.935)	0.672(0.062)	0.261(0.095)	0.345(0.121)	0.424(0.136)
Iris $R = 50$	8(0.953)	0.640(0.069)	0.313(0.069)	0.402(0.125)	0.454(0.137)

FP(False Positives - lower the better), TP_r=Sens= $\frac{\#TP}{\#TP+\#FN}$ (Detection Rate/Sensibility - higher the better). AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better).

Is possible to observe that the number of FP reduced drastically in all methods and the Tp_r decrease in small proportion due to some True Positives (TP) being discarded. Region metrics also decreased in a small proportion due to the rejection of some positives cases. In watershead methods the nipples were detected as masses in a vast number of occasions since it present high-density tissue and contour similarities, misleading the detection and posterior FP reduction. Previous nipple pre-processing should be addressed to tackle this problematic.

3.6 Calcification Lesion Detection

Calcifications are characterized for being small and bright. These characteristics require different approaches for its detection and characterization. We describe two main detection methods, based on outlier detection and top hat filtering.

3.6.1 Outlier Detection

The task of automatically detect calcifications can be seen as identifying pixel values that greatly differ from the normal image intensity distribution. A definition of Inter Quartile Range (IQR) can be expressed as

$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR] \tag{3.38}$$

where Q_1, Q_3 correspond to the 1st and 3rd quartile respectively and IQR the range between both as represneted in Figure 3.26.

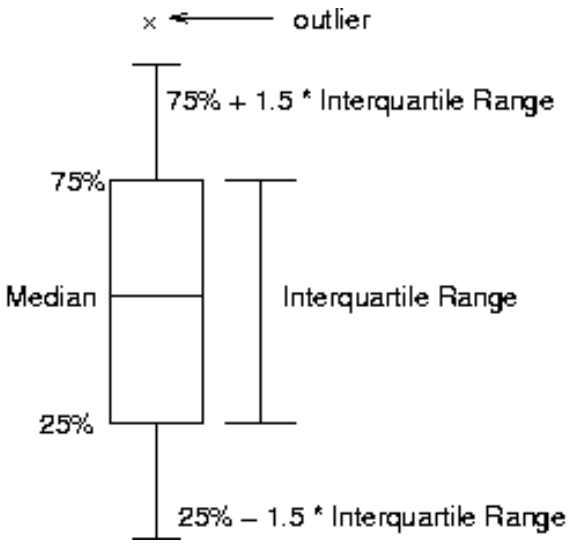
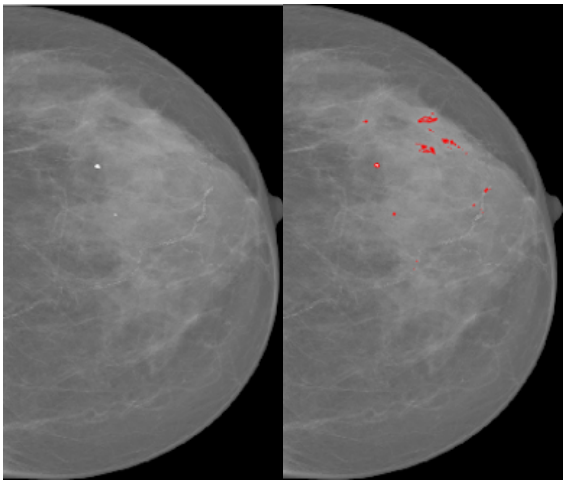


Figure 3.26: Boxplot.

Intensity values above the upper range can be seen as pixels that differ so much from other values that arouse suspicious that were generated by a different mechanism (Figure 3.27).



Original 1 Outlier 1

Figure 3.27: Outlier detection (Red - Detection's).

3.6.2 2D Wavelet Decomposition

Considering the 1D Fourier Transform (Chui, 2016) defined by $exp(jwt)$ were t corresponds to the time domain that defined the function of time to be converted into a function of frequency w. For the 2D case, the Fourier Transform can be obtained as

$$\exp(j(w_1 t_1 p w_2 t_2)) \tag{3.39}$$

The transformed coefficient becomes two variable functions so as the 2D discrete wavelet transform (Salve and Chakkarwar, 2013). Denoting the scaling and wavelet function as $\phi(x, y)$ and $\psi(x, y)$, the scaled and translated basis functions can be defined as:

$$\begin{aligned} \phi_{j,m,n}(x, y) &= j^{j/2}\phi(2^j x - m, 2^j y - n), \\ \psi^i_{j,m,n}(x, y) &= j^{j/2}\psi^i(2^j x - m, 2^j y - n), \quad i = \{H, D, V\} \end{aligned} \tag{3.40}$$

where $x$and$y$ are pixels indices and $m$and$n$ horizontal and vertical stride, three different wavelet functions $\psi^H(x, y)$, $\psi^V(x, y)$ and $\psi^D(x, y)$ can be defined. Conceptually, the scaling function is the low-frequency component of the previous scaling function in 2 dimensions. Therefore, there is one 2D scaling function. However, the wavelet function is related to the order to apply the filters. If the wavelet function is separable, i.e. $f(x, y) = f_1(x)f_2(y)$ these functions can be easily rewritten as

$$\begin{aligned} \phi(x, y) &= \phi(x)\phi(y), \\ \psi^H(x, y) &= \psi(x)\phi(y), \\ \psi^V(x, y) &= \phi(x)\psi(y), \\ \psi^D(x, y) &= \psi(x)\psi(y), \end{aligned}$$

Defining the functions as separable simplifies its analysis the 2D function enabling to focus on the design of 1D wavelet and scaling functions. The analysis and synthesis equations are modified to

$$\begin{aligned} W_\phi(j_0, m, n) &= \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)\phi_{j_0,m,n}(x, y) \\ W^i_\phi(j_0, m, n) &= \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)\phi^i_{j_0,m,n}(x, y), i = \{H, V, D\}, \\ f(x, y) &= \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} W_\phi(j_0, m, n)\phi_{j_0,m,n}(x, y) \\ &+ \frac{1}{\sqrt{MN}} \sum_{i=H,V,D} \sum_{j=j_o}^{\inf} \sum_m \sum_n W^i_\psi(j, m, n)\phi^i_{j_0,m,n}(x, y) \end{aligned}$$

This is the general form of a 2D wavelet transform. If the scaling and wavelet functions are separable, the summation can be decomposed into two stages. The first step is along the x-axis and then calculate along the y-axis. For each axis, a fast wavelet transform can be applied to accelerate the speed. A schematic diagram is shown in Figure 3.28.

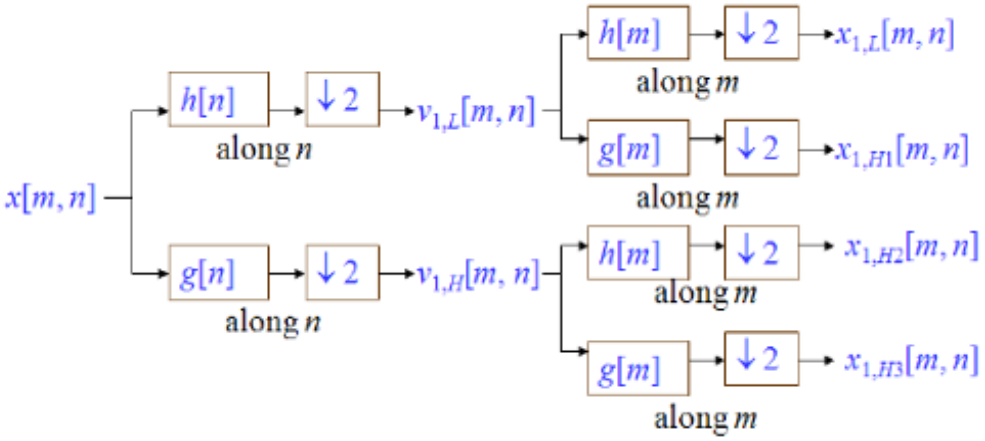


Figure 3.28: Schematic diagram of 2D wavelet transform.

This kind of two-dimensional Discrete Wavelet Transform (DWT) leads to a decomposition of approximation coefficients at level j in four components: the approximation at level $j+1$, and the details in three orientations (horizontal, vertical, and diagonal). The two dimensional signal (usually image) can be observed on Figure 3.29 showing the decomposition of a image into the four bands: LL(left-top), HL(right-top), LH(leftbottom) and HH(right-bottom). The HL band indicates the variation along the x-axis while the LH band shows the y-axis variation.

Original Wavelet decomposition

Figure 3.29: Lena image before and after wavelet decomposition.

In the point of coding, we more bits can be spent on the low-frequency band and less bit on the high-frequency band or even set them to zero if we want to remove high-frequency components.

3.6.3 Experiments and Results for Detection of Suspicious Calcification Lesions

For the detection of calcification's two approaches were implemented, first based in outlier pixels detection and the second corresponding to a combination of Top hat filtering followed by a wavelet decomposition to attain high-frequency components. In Top Hat filtering the selected structure elements corresponds to 8 different set of lines revolving around the center of a 9×9 pixel array, combined with a subtraction between the original image and the maximum of the opening results to obtain an image. For the wavelet decomposition, the choice was the second-order symmetrical wavelets as the decomposition filters since they have the least asymmetry and highest number of vanishing moment (Daubechies, 2016). Figure 3.30 present side by side comparison between the two methods.

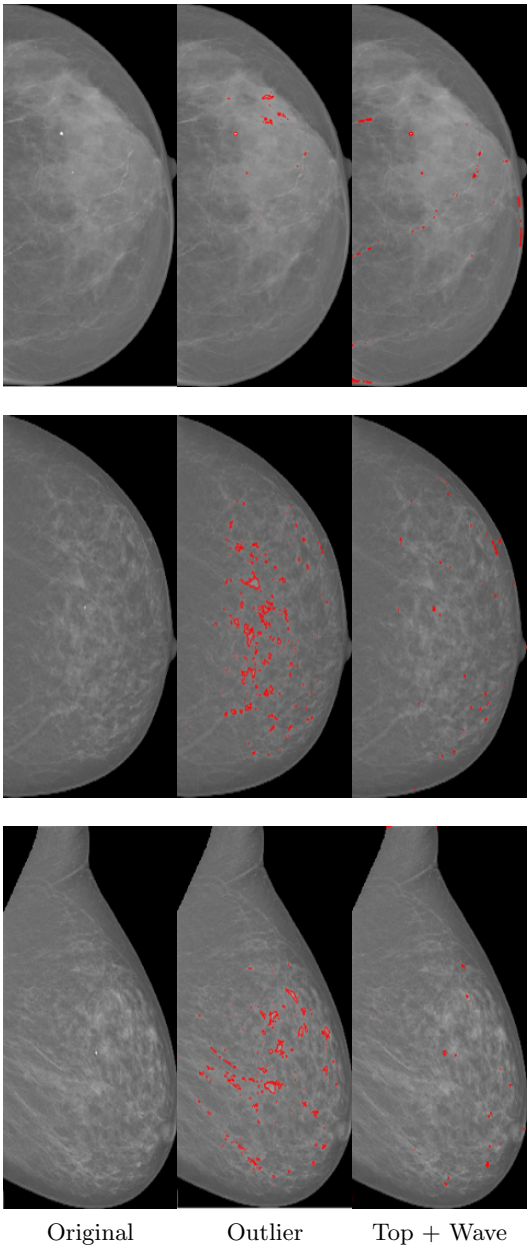


Figure 3.30: Example of the calcification detection (Blue - GT and Red - Detection).

Table 3.8 summarizes the results for calcification detection. For base comparison, the method proposed by Trovini et al. (2018) using CNN trained in a patch-based fashion on INbreast and private dataset to detect calcification's, yielding and Area Under ROC curve (AUC) of 0.9998.

Table 3.8: Performance evaluation for detection of suspicious calficication lesions . Results mean (std).

Method	FP	TP_r	AUC
SotA.	-	-	0.9998(-)
Outlier	58(11.012)	0.411(0.011)	-
Morp + Wav	47(9.045)	0.326(0.092)	-

FP (False Positives - lower the better), TP_r=Sens= $\frac{\#TP}{\#TP+\#FN}$ (Detection Rate/Sensibility - higher the better). AUC (Area Under Curve - higher the better), Measures range from $[0, 1]$

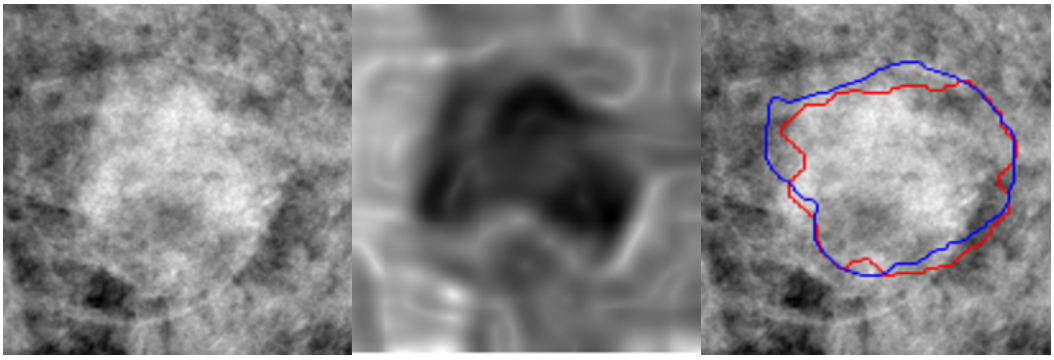
Outlier detection was able to detect a some of calcification's $TP_r \approx 0.411$, however at the cost of a high number of FP. The morphological operation, namely Top Hat transformation followed by wavelet decomposition attaining only high-frequency components achieve a lower number of FP, however with lower TP_r.

3.7 Mass Lesion Contour Extraction

After mass detection, the next task consists in extracting mass contour to visually characterize the mass genre and assess the severity of the anomaly. For the task, Snakes, SP in polar and Cartesian coordinates, and Sliding band Filters (SBF) were evaluated against state of the art methods and baseline.

3.7.1 Snake Segmentation

Snakes segmentation, described in 3.4.3 can be employed to the task of extracting the mass contour. An initial snake is set as a circular ellipse contained in the detection bounding box, that thought the iterations of the algorithm will be fitted to the mass external contour (Figure 3.31).



Original Snake Energy Final

Figure 3.31: An example of the snake contour extraction (Blue - GT, Red - Detections)

3.7.2 Shortest Path in Polar Coordinates (SPPC)

SP in Polar Coordinates algorithm (SPPC) described in 3.4.5 can be tasked to obtain the mass contour. The main difference regarding the pectoral muscle segmentation relies upon the fact that the transformation is carried out over the center coordinates of the mass region instead of the top left image point. Additionally, an extra constraint is added to force the endpoint to be coincident with the starting point. After the minimal path is found, the image is transformed back to the original coordinate system and the final contour is attained.

3.7.3 Shortest Path in Cartesian Coordinates (SPCC)

The computation of the closed shortest path in the original coordinates has some initial difficulties. To address this difficulties, Cardoso et al. (2015) starts by creating a Directed Acyclic Graph (DAG) from the grid and respective linearization. Secondly, since paths closer to the center have fewer pixels, they will be naturally selected even if the cost of the weight of the edges is slightly lower than in paths close to the center. Thirdly, one must consider if the Euclidean distance between nodes (pixels) is appropriate or if another notion of distance is more effective to capture the distance in the context of closed paths enclosing a given node. While working in polar coordinates, the linear ordering of the vertices is given by column number since one wants to go from left to right or vice-versa. Working in the original coordinate space, one wants to go around a given seed point ϱ ranging from $[0, 360]$ degrees. Therefore is natural to order nodes by the angle θ of the node relative to ϱ. Causal neighbors of a given node correspond to neighbors with lower q with edges being oriented from the causal neighbors towards the seed point. The number of causal neighbors varies from node to node, depending on q and the given position of the node in the ROI. Edges are oriented from the causal neighbors to the given point (Figure 3.32).

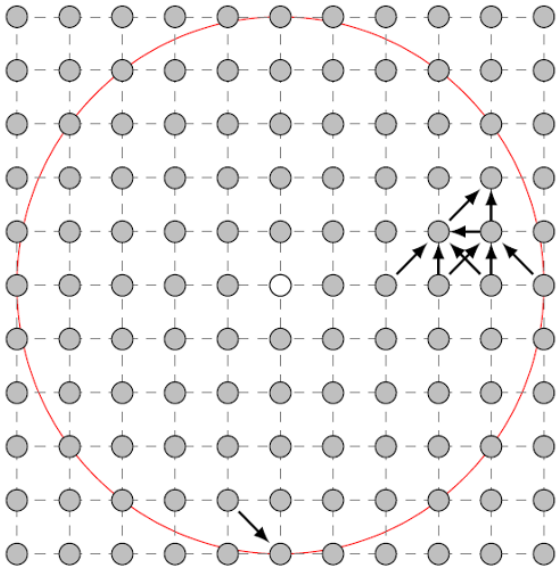


Figure 3.32: For a ROI with a radius of 5 (red) and an 8-neighbourhood, this figure illustrates the causal neighbours for a few nodes. The number of causal neighbours varies from 1 to 4. (Image from Cardoso et al. (2015))

On Figure 3.33, two closed paths enclosing a point C are presented, with pixels assuming same value on both contours. With the cost of the edge relying solely on the features extracted from edge or neighborhood of the pixel, both contours have the same cost and the smaller contour is selected since it has fewer edges, therefore smaller overall cost. To circumvent the problem of small paths collapsing from a given point C being naturally favored, the edge cost is manipulated to adapt to the correct increase of the number of edges in a path with a given distance. The cost of an edge is now weighted by $1/r$, where r is the distance of the head node to the edge, resulting in the perimeter (and therefore, the number of edges in the contour) growing proportionally to r, making the overhaul cost approximately independent of r.

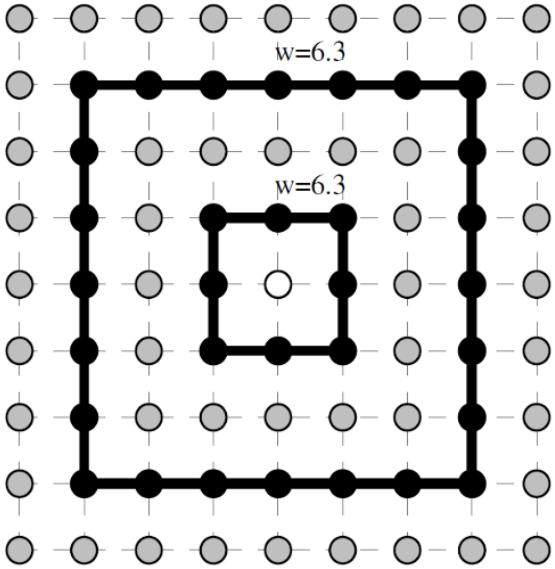


Figure 3.33: Two closed paths enclosing the centre of the ROI. Without a proper modulation, the inner path presents a smaller overall cost. (Image from Cardoso et al. (2015))

Regarding the distance between nodes, the cost of an edge includes a factor related with the feature(s) computed at the head of the edge and a factor related with the distance between the head and the tail of the edge. Focusing on the second the Euclidean distance between the two nodes appears as a reasonable solution. In this case, the position of the seed point does not impact the distance (Figure 3.34). In one situation the contour is moving directly to the center, while in the other, the contour is almost orthogonal to the radius of the current node. In the polar coordinates transformation, not only all circular paths on the center of the ROI have the same cost but they are also the shortest paths when all the pixels have the same information. In fact, they are transformed into straight lines between opposite margins

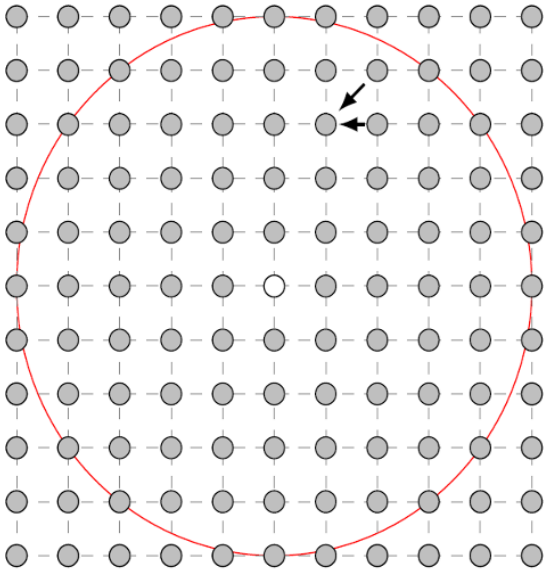


Figure 3.34: Two movements with different characteristics. (Image from Cardoso et al. (2015))

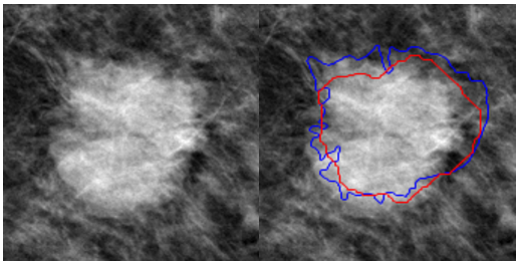
To mimic this behaviour two other measures can be used, with the first corresponding exactly from what would be obtained in the polar domain, using a resolution of one degree per pixel and radius unit per pixel

$$d_{polar} = \sqrt{(\Delta r)^2 + (\Delta \theta)^2} \tag{3.41}$$

The second measure can correspond to the Euclidean distance modulated by a function of α, corresponding to the angle between the orthogonal direction of the radius at current node and the vector from the casual neighbor to the current point as

$$d_{cos} = \frac{d_{Euclidean}}{\cos \alpha} \tag{3.42}$$

Figure 3.35 shows mass lesion patch contour segmentation using the 8-neighborhood and polar distance. The weight was set as a nonlinear function of the derivative.



Mass 1 Contour 1

Figure 3.35: Mass examples (Red - Detection's, Blue - GT).

3.7.4 Sliding Band Filter (SBF)

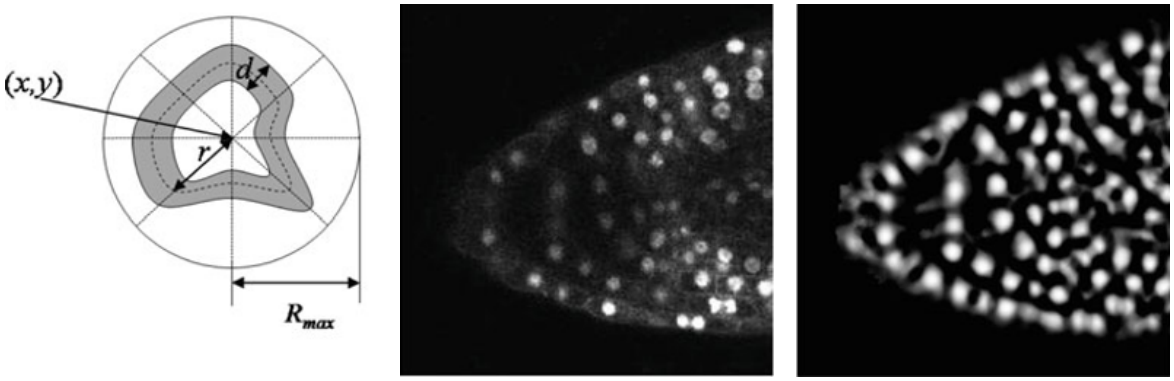
Sliding Band Filter (SBF) (Esteves et al., 2012) combines the ideas of Iris Filter (IF) (Kobatake and Hashimoto, 1999) and Adaptative Ring Filter (ARF), (Wei et al., 1999) by defining a support region formed by a fixed width band, with varying radius in each direction, allowing maximization of the convergence index at each point, (Pereira et al., 2007). The SBF formulation can be derived from ARF and IF convergence estimation as

$$SBF(x,y) = \frac{1}{N}\sum_{i=0}^{N-1} \max_{R_{min}\leq r\leq R_{max}} \left[\frac{1}{d}\sum_{m=r-d/2}^{r+d/2} CI(,x,y,i,m)\right], \tag{3.43}$$

where d corresponds to the width of the band, moved between R_{min} and R_{max}. The shape estimation of the SBF is similar to the IF, (Kobatake and Hashimoto, 1999). The corresponding shape radius for each radial line is given as

$$r_{shape}(x,y,i) = argmax_{R_{min}\leq r\leq E_{max}} \left[\frac{1}{r}\sum_{m=r-d/2}^{r+d/2} CI(i,m)\right] \tag{3.44}$$

This filter combines both the shape flexibility of the IF with the limited band search of the ARF. The resulting estimated shapes are similar to those obtained using the IRIS filter (IF) with respect to shape ranges.



(a) SBF filter (b) Original (c) SBF responses

Figure 3.36: Schematic of the filter support region of the SBF filter (Support region as grey), Original Image and SBF responses.

3.7.5 SBF with Phase Congruence

While image magnitude derivatives are poorly defined in low contrast areas, image phase information can be used to extract meaningful information on edge direction and strength, (Kube, 1992). Phase information provides greater robustness in low contrast, motivating the development of phase-based edge measures. Phase Congruence or Coherence (PC) attempt to find locations in an image where all sinusoids in the frequency domain are in phase. These locations generally correspond to the location of a perceived edge regardless of whether the edge is represented by a large or small change in intensity on the spatial domain.

PC is a directional measure evaluated over a range of orientations and the resulting edge evidence image result corresponds to the sum of all individual responses. PC can be defined by making use of the image wavelet transform. Given an image $I(x, y)$ and the even-symmetric (cosine) and odd-symmetric (sine) wavelet signals $M_n j^e$ and $M_n j^o$ at scale n, respectively, a wavelet transform of the image can be obtained as

$$[e_{nj}(x, y), o_{nj}(x, y)] = [I(x, y) * M^e_{nj}, I(x, y) * M^o_{nj}], \tag{3.45}$$

where $*$ is the convolution operation and j the orientation under analysis. Given wavelet responses $[e_{nj}(x, y), o_{nj}(x, y)]$ for scale n and orientation j a response amplitude and phase is defined as:

$$\begin{aligned} A_{nj}(x, y) &= \sqrt{e_{nj}(x, y)^2 + 0_{nj}(x, y)^2}, \\ \phi_{nj}(x, y) &= \tan^{-1}\left(\frac{e_{nj}(x, y)}{o_{nj}(x, y)}\right) \end{aligned} \tag{3.46}$$

A phase congruence for specific orientation j based on $A_{nj}(x, y)$ and $\phi_{nj}(x, y)$ can be defined

as:

$$PC(x,y,j) = \frac{\sum_n W_j(x,y)\lfloor A_{nj}(x,y)\Delta\Phi_{nj}(x,y) - T_j\rfloor}{\sum_n A_{nj}(x,y) + \epsilon} \tag{3.47}$$

where $\lfloor\rfloor$ is the floor round, setting the enclosed quantity equal to itself if positive or zero otherwise, ϵ is a small positive constant to avoid divisions by zero on locations where the wavelet response approaches zero, T_j is the noise estimate based on high-frequency wavelet responses, $W_j(x,y)$ is a weighting function that penalizes filter response spread and $\Delta\Phi_{nj}(x,y)$ corresponds to a sensitive phase deviation function defined as:

$$\begin{aligned}\Delta\Phi_{nj}(x,y) = \cos(\phi_{nj}(x,y) - \bar{\phi_j}(x,y) \\ -|\sin(\phi_{nj}(x,y) - \bar{\phi_j}(x,y)|\end{aligned} \tag{3.48}$$

where $\bar{\phi_j}$ is the average phase for location (x,y) on orientation j.

PC is a measure that is symmetric between $[0,\pi]$ and $[\pi,2\pi]$ as such only PC estimates for half of the filters radial directions are required. So is possible to re-write Equation 3.47 as

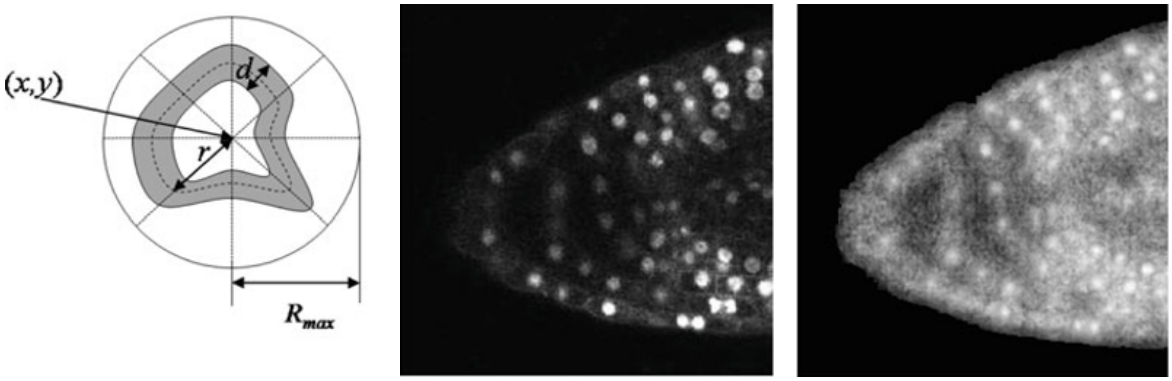
$$\begin{aligned}SBF_{PC}(x,y) = \frac{1}{N}\sum_{i=0}^{N-1}\max_{R_{min}\le r\le R_{max}} \\ \times\left[\frac{1}{d}\sum_{m=r-d/2}^{r+d/2} PC(x_i,y_i,j(i))\right]\end{aligned} \tag{3.49}$$

with $x_i = x + m*\sin(i)$, $y_i = y + m*\cos(i)$ and $j(i) = \begin{cases} i, & i \le N/2 \\ i - \frac{N}{2} & i > N/2 \end{cases}$

3.7.6 SBF Filter with Shape Regularization

SBF filters tend to better separate overlapping regions if they exist. A final shape regularization, namely a radial active contour fitting is introduced by Esteves et al. (2012), combining radial shape smoothness and image energy based on gradient convergence to refine the final mass contour. The total energy that governs the mass shape regularization process is defined by

$$E_{total} = E_{internal}(r_{shape}) + \gamma E_{external}(r_{shape}, CI) \tag{3.50}$$



(a) SBF filter (b) Original (c) SBF responses

Figure 3.37: Schematic of the filter support region of the SBF filter (Support region as grey), Original Image and SBF Phase responses.

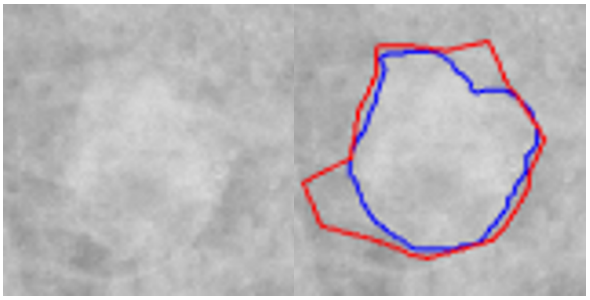
where γ controls the degree of regularization, e_{shape} is a variable containing information about the support point locations defined in Equation 3.44 and CI corresponds to the Convergence Index (Equation 3.34). Shape smoothness energy for the mass shape $E_{internal}(r_{shape}$ can be defined as:

$$E_{internal}(r_{shape}) = \alpha \left| \frac{\delta r_{shape}(i)}{\delta i} \right|^2 + \beta \left| \frac{\delta r_{shape}(i)}{\delta i} \right|^2, \tag{3.51}$$

where α and β control the degree of elasticity and stiffness respectively while i corresponds to the radial index in the specific radial distance of the shape r_{shape}. The image convergence energy that regulates the fitting of final shape to the underlying image information is given by

$$E_{external}(r_{shape}, CI) = \frac{1}{N} \sum_{i=1}^{N} CI(x, y, i, r_{shape}(i)), \tag{3.52}$$

with CI as the convergence index for the shape described by the radial distances contained in r_{shape} at position (x, y). Figure 3.38 present extracted contours with shape regularization.



Mass 1 Contour 1

Figure 3.38: Two mass examples (Red - Detection's, Blue - GT).

3.7.7 Experiments and Results for Mass Lesion Contour Extraction

For mass lesion contour segmentation, 4 methods were implemented and evaluated. Each image patch in increased by 20% to attain the surrounding lesion area.

The SP in Cartesian Coordinates the cost function was set to be the the radial image derivative combined with a an exponential law for weight creation expressed as

$$\hat{f}(g) = f_l + (f_h - f_l)\frac{\exp\left((255 - g)\cdot\beta\right) - 1}{\exp\left(255\cdot\beta\right) - 1} \tag{3.53}$$

with $f_h, f_l, \beta \in \mathbb{R}$ set to constants values $f_h = 30, f_l = 2, \beta = 0.025$ and g the minimum of the gradient on the two incident pixels.

For SP in Polar Coordinates, the center of polar transformation was set to be the patch image center, $f_h, f_l, \beta \in \mathbb{R}$ set to values $f_h = 25, f_l = 4, \beta = 0.022$ respectively. Directional cost parameters $C_{right} = 2.2$ and $C_{left} = 1.5$ with additional cost manipulation to force the initial and final points to be on the same location.

For snake method, we set the number of interactions $Iter = 3000$, the attractiveness towards black lines $w_{line} = -0.29$, the edge attractiveness $W_{edge} = 3$, attraction to termination lines $E_{term} = 0.05$ and $\sigma = 5$ used to calculate the gradient of edge energy.

For Sliding Band Filters with Shape Regularization proposed by Esteves et al. (2012) the radius parameters were set to $R_{min} = 4$, $R_{max} = 70$ to attain small and larger lesions, the number of orientations $N = 36$, the Gaussian smoothing kernel set with $\sigma = 1$, and internal and external contour snake energy set to $reg_a = 2$ and $reb_b = 2$ respectively, enabling to capture outside lesion contour.

In addition, a baseline segmentation method is included consisting in the circular perimeter that encompasses the exterior mass contour. As state of the art base comparison, Cordeiro et al. (2016) uses GrowCut technique were the user initially labels a set of pixels in different

classes of interest and, based on these seeds, the algorithm tries to label all the pixels of the image. It yielded a DICE coefficient of 0.823(0.046) in INbreast database.

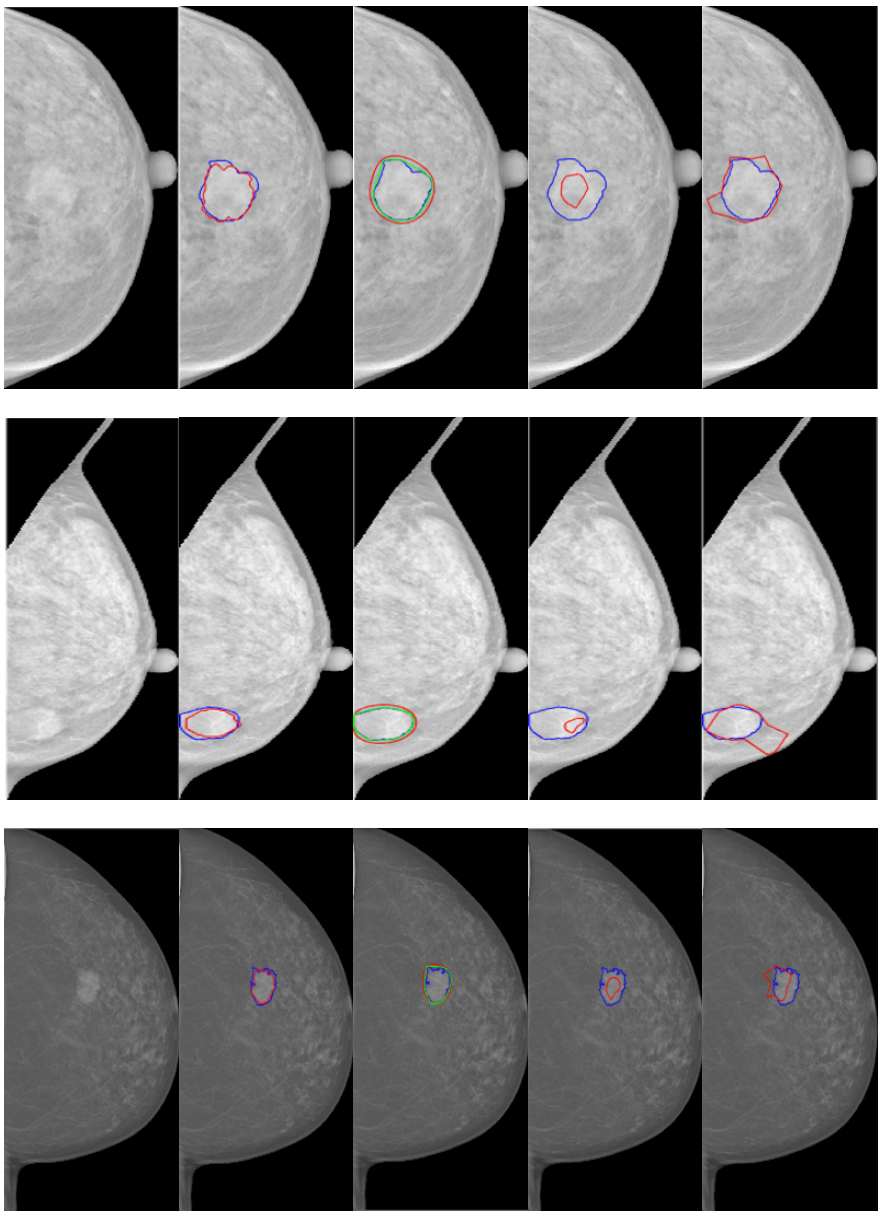


Figure 3.39: Example of the mass contour extraction (Blue - GT and Red - Detection).

The Table 3.5 presents the computed metrics for the mass contour detection, namely Area Overlap Measure (AOM), Combined Measure (CM) and DC. The results are the mean (std).

Table 3.9: Performance evaluation of mass contour extraction. Results are in mean (std).

Method	AD	AMED	HD	AOM	CM	DICE
SotA	-	-	-	-	-	0.823(0.046)
Baseline	10.001(5.623)	10.572(6.052)	26.632(15.503)	0.746(0.132)	0.751(0.088)	0.762(0.105)
SPCC	6.824(7.719)	7.655(8.289)	22.148(19.354)	0.792(0.108)	0.836(0.103)	0.841(0.107)
Snakes	20.753(14.504)	7.093(4.721)	17.409(11.086)	0.712(0.132)	0.781(0.117)	0.743(0.122)
SPPC	20.753(14.504)	24.644(17.179)	43.086(28.230)	0.743(0.101)	0.750(0.106)	0.803(0.115)
SBF-Reg	17.272(10.166)	20.687(11.900)	53.146(24.754)	0.7017(0.107)	0.591(0.128)	0.604(0.123)

AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better), while AD, AMED and HD are measures of pixel error (the lower the better).

Considering the accuracy metrics (AOM, CM and DICE), the SP in Cartesian coordinates exhibit the better performance followed by SP in polar coordinates. This difference is due to the fact that no transformation of the image is done contrary to the SP in polar coordinates. In addition, SBF filters with regularization and snakes exhibited a reasonable performance, however lower than the baseline due to over-segmentation in many cases. SBF also has a higher computational cost. Compared to the proposed state of the art method, all the evaluated methods don't require any manual user input, making suitable for building an automatic CAD system.

3.8 Summary

Pectoral muscle segmentation can be used to refine the data to use as input for the consent stage maximizing the expected lesion segmentation detection quality. U-Net segmentation has some possible enhancements that might improve the obtained results, such as extended data augmentation through the introduction of enhancement processes that further refinement the segmentation process and different output metrics that capture other segmentation semantics.

Concerning lesion contour extraction, extensive parameter tuning can improve snake and SBF methods to be closer to SP in Cartesian Coordinates.

Chapter 4

Breast Structures Classification

The majority of real-world classification problems require supervised learning where the underlying class probabilities and class-conditional probabilities are unknown and each instance is associated with a class label. Improper features can degrade the performance of the models, by making then unable to generalize very well, compromising its reliability. Section 4.1 details the employed methods for feature analysis, dimensionality reduction and selection, and Section 4.2 presents a brief description of the classification models and main components complemented with description of Convolutional Neural Networks (CNN) models for image screening and patch classification. Each of the identified components is complemented with a description of the conducted experiments discussion of the attained results.

4.1 Feature Analysis and Selection

Feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building be better generalizable models, and decreasing required storage. For the classification problem, feature selection aims to select a subset of highly discriminant features. In other words, it selects features that are capable of discriminating samples that belong to different classes. Relevant feature is neither irrelevant nor redundant to the target concept; an irrelevant feature is not directly associate with the target concept but affect the learning process, and a redundant feature does not add anything new to the target concept (Dash and Liu, 1997). In many classification problems,

it is difficult to learn good classifiers before removing these unwanted features due to the huge size of the data. Reducing the number of irrelevant/redundant features can drastically reduce the running time of the learning algorithms and yielding a more general classifier. In this section, we present the features analysis in term of feature dimensionality, information gain and correlation to support the final selection of features.

4.1.1 Information Gain

Due to its computational efficiency and simple interpretation, information gain is one of the most popular feature selection methods. It is used to measure the dependence between features and labels by calculating the information gain between the i-th feature f_i and the class labels C as

$$IG(f_i, C) = H(f_i) - H(f_i|C) \tag{4.1}$$

where $H(f_i)$ is the entropy of f_i and $H(f_i|C)$ is the entropy of f_i given C

$$\begin{aligned} H(f_i) &= -\sum_j p(x_j)\log_2(p(x_j)), \\ H(f_i|C) &= -\sum_k p(c_k)\sum_j p(x_j|c_k)\log_2(p(x_j|c_k)) \end{aligned} \tag{4.2}$$

In information gain, a feature is relevant if it has a high information gain. Features are selected in a univariate way, therefore, information gain cannot handle redundant features.

4.1.2 Principal Component Analysis and Bi-Plots for Mass Lesion Feature Analysis

Principal Component Analysis (PCA) is a standard statistical technique that can be used to reduce the dimensionality of a data set. Has proven to be an exceedingly useful tool for dimensionality reduction of multivariate data. PCA transforms the initial data set represented by vector samples into a new set of vector samples with derived dimensions. The basic idea implies that a vector sample $x = \{x_1, x_2, \cdots, x_n\}$ should be transformed into a set $Y = \{y_1, y_2, \cdots, y_n$ with the same dimensionality, but with Y having different properties, with most of the information being held in the first dimensions. As result, Figure 4.1 with the first row describing the standard deviation associated with each Principal Component and the second row showing the proportion of the variance in the data explained by each component and the curve describes the cumulative proportion of explained variance.

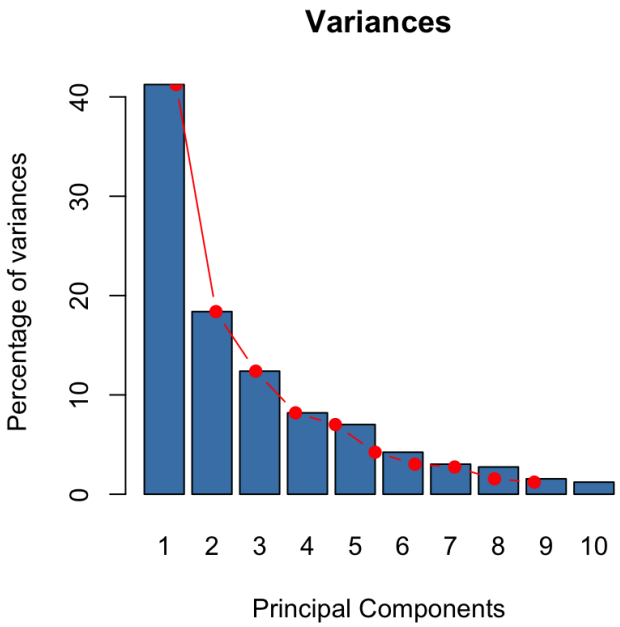


Figure 4.1: Plot of percentage of explained variance versus dimension considered on shape features (Image from [6])

with the first five Principal Components accounting for more than 95% of the variance of the data.

Bi-plots is a type of exploratory graph used in statistics, corresponding to a generalization of the simple two-variable scatter-plot. It allows information on both samples and variables of a data matrix to be displayed graphically. Samples are displayed as points while variables are displayed either as vectors (Figure 4.2). A biplot is a useful tool for visualizing the results of PCA. It allows you to visualize the principal component scores and directions simultaneously.

[6] http://www.sthda.com/english/wiki/print.php?id=207

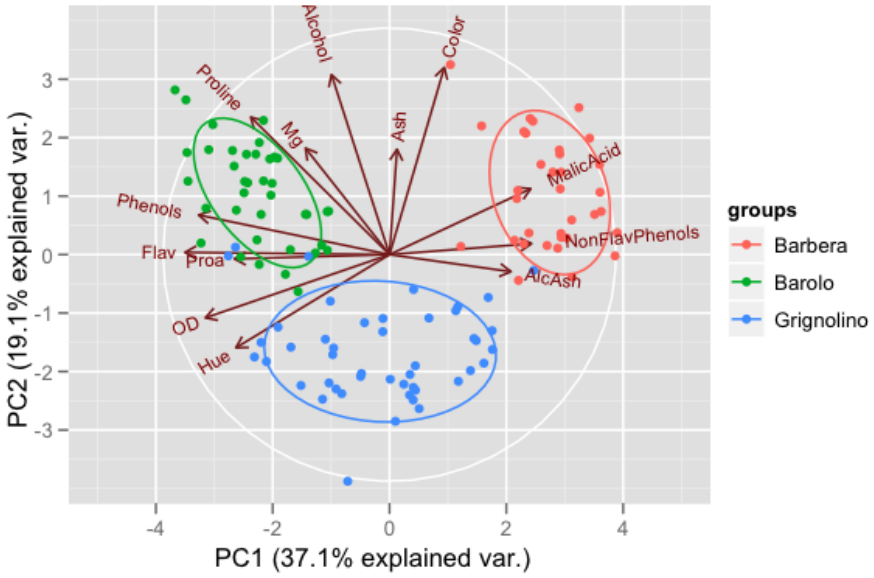


Figure 4.2: Bi-plot diagram of the shape features (Image from [7]).

4.1.3 Feature Selection

In order to test which features are informative, three different metrics can be used, namely the Pearson correlation, the distance correlation and the Maximal information coefficient.

The Pearson correlation (*corr*) is the most commonly used measure and quantifies the linear dependence between two variables. *corr* can assume values between -1 and +1, inclusive, where +1 corresponds to total positive correlation, 0 is no correlation, and -1 is a total negative correlation.

$$corr = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2]\,[N\sum Y^2 - (\sum Y)^2)]}} \tag{4.3}$$

where N corresponds to the number of pair scores, $\sum XY$ the sum of the product of paired scores, $\sum X$ and $\sum Y$ cum of X and Y scores respectively and $\sum X^2$, $\sum Y^2$ the sum of squared of X and Y respectively.

The distance correlation (*dcorr*) (Székely and Rizzo, 2009) characterizes independence: it is zero if and only if the vectors are independent. Comparing with the Pearson correlation, *dcorr* measures not only linear associations but all types of dependence relations. The distance correlation is comprehended between $0 \leq dcorr \leq 1$.

[7] https://stats.stackexchange.com/questions/7860/visualizing-a-million-pca-edition

4.1.4 Experiments and Results for Feature Analysis and Selection

Considering the set of features described in Table 2.5 in addition with Difference Center of Weighted Center (dCwC), a detailed analysis was performed using PCA. All the extracted features were normalized with zero mean and unit variance. The experiments presented were made with the INbreast images containing masses only (without calcification's) and excluding examples with Breast Imaging Reporting And Data System (BI-RADS) class 1.

4.1.4.1 PCA Feature Analysis

Considering the PCA analysis (Figure 4.3) is possible to verify that the first five Principal Components accounts for more than 95% of the variance of the data. Considering that PCA accounts for variance of each feature among corresponding classes, is assumed that features that present high variance are more likely to have a good split between classes.

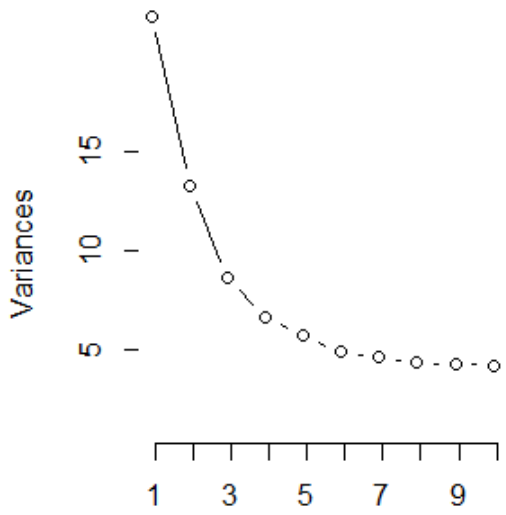


Figure 4.3: Plot of percentage of explained variance versus dimension considered features.

4.1.4.2 Information Gain Feature Analysis

In what concerns information gain, two methods were used to access feature importance: (1) mean decrease impurity and (2) mean decrease accuracy.

(1) *Mean decrease impurity* accounts for (locally) optimal conditions, defined as the im-

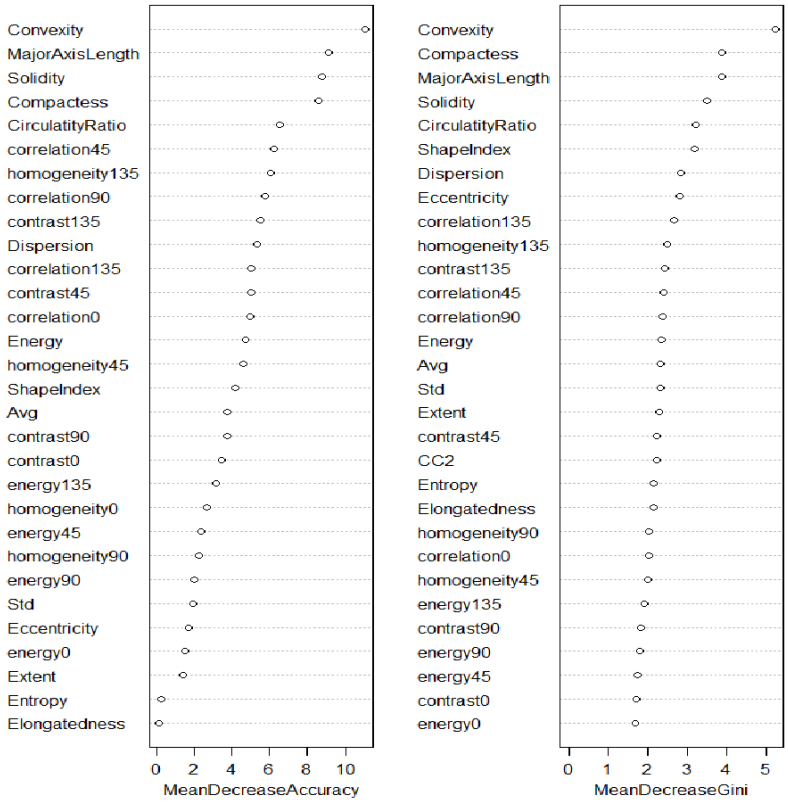


Figure 4.4: Plot of feature importance (only most important)

purity, that on classification tasks can be typically either Gini impurity or information gain/entropy.

On the other hand, (2) on *Mean decrease accuracy* the general idea is to permute the values of each feature and quantify how much this permutation decreases the accuracy of the model. Unimportant variables should have little to no effect on model accuracy when permuted.

Figure 4.4 presents the rank of the feature importance.

The results enable to conclude the existence of several variables are not very relevant and need to be properly addressed.

4.1.4.3 Correlation Analysis

For the final feature selection, in order to test which features are informative, two different metrics were used, the Pearson and the Distance correlations.

To access which were the most informative features, the correlation between each feature

vector and the corresponding BI-RADS class was computed. The hypothesis of no correlation against the alternative (non-zero correlation) was tested. Two significance values of 0.1 and 0.05 were considered. Features that present *pvalues* smaller than the defined significance were kept, summarized in Tables 4.1 and 4.2.

Table 4.1: Selected mass features ($pvalue < 0.1$).

Feature	Acron	*corr*	*dcorr*
Aspect ratio	Asp	0.451	0.578
Circularity	Circ	0.151	0.063
Compactness	Com	0.342	0.327
Contained lines	Cl	0.563	0.627
Convexity fraction	fCV	0.582	0.604
$CC_2 = \sqrt{\frac{r_{min}}{R_{max}}}$	CC_2	0.401	0.414
Difference Center and Weighted Center	dCwC	0.279	0.236
Eccentricity	ECT	0.134	0.116
Entropy Radial Length Histo	ERLH	0.534	0.614
Histograms of Gradient Divergence	HGD	0.739	0.751
Lobulation Index	LI	0.462	0.521
Perimeter	Per	0.154	0.050
Roundness	RND	0.234	0.313
Shape Index	ShI	0.396	0.548
Sharpness Index	Sh	0.635	0.673
Skeleton end points	SEP	0.575	0.587
Solidity	Sol	0.632	0.674
Spiculation	Sp	0.734	0.746

cor and *dcorr* are measures of correlation ranging from $[0, 1]$.

Table 4.2: Selected mass features ($pvalue < 0.05$).

Feature	Acron	*corr*	*dcorr*
Aspect ratio	Asp	0.451	0.578
Compactness	Com	0.342	0.327
Contained lines	Cl	0.563	0.627
Convexity fraction	fCV	0.582	0.604
$CC_2 = \sqrt{\frac{r_{min}}{R_{max}}}$	CC_2	0.401	0.414
Difference Center and Weighted Center	dCwC	0.279	0.236
Eccentricity	ECT	0.134	0.116
Entropy Radial Length Histo	ERLH	0.534	0.614
Histograms of Gradient Divergence	HGD	0.739	0.751
Lobulation Index	LI	0.462	0.521
Roundness	RND	0.234	0.313
Shape Index	ShI	0.396	0.548
Sharpness Index	Sh	0.635	0.673
Skeleton end points	SEP	0.575	0.587
Solidity	Sol	0.632	0.674
Spiculation	Sp	0.734	0.746

cor and *dcorr* are measures of correlation ranging from $[0, 1]$.

From both tables is possible to conclude that features that describe contour irregularity were kept, presenting also high correlation with the corresponding BI-RADS class, reinforcing the initial considerations that the higher malignity are associated with irregular contours. Also, the use of a lower significance value results in the remove of two features, Circularity and Perimeter that have a lower correlation with the BI-RADS class since most of the mass lesions present circular shape with small circularity variation in malignant cases. For final feature selection, the significance threshold *pvalue* was set to 0.1 and features bellow were thus kept.

4.2 Mass Lesion Classification

Classification of mass lesions is vital to asses the degree of severity in an automatic way. For the task, models make use of extracted features from lesions to learn models able to infer about the BI-RADS class. Several models, combinations together with their best parameters can be evaluated with the objective to obtain the best model. The classification can be binary (benign/malign) or multi-class.

4.2.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed both for classification and regression task. SVMs base idea relies upon finding the hyperplane that best divides two district classes. SVM was initial proposed by Cortes and Vapnik (1995) as a binary classifier. SVM starts by separating the feature space through a hyperplane and finding the maximal separating line using only two class points (Support Vectors) that lay near the frontier line. The separating hyperplane is defined by orientation and distance to the original plane, with parameters being optimized by quadratic programming algorithms or gradient descent based methods. SVM can be extended to multi-class classification by employing one-vs-all approaches, were the separation hyperplane for a specific class is optimized considering the selected class versus all the remaining classes. This procedure is performed for all classes, achieving a multi-class classifier.

Non-linear data can be also separated by SVM by using the kernel methods that handle the curse of dimensionality, by avoiding the explicit mapping of data into a high dimensional space. Kernel methods start by transform the input variable space to an implicit feature space where can be linear separable, but without ever computing the coordinates of the data in that space. The kernel function,$k(x; x_0)$, is used to simplify the computation of inner products between all pairs of input variables in the original space (Bishop, 2006). Formally, given the training set $\{x_i, y_i\}_{i=1}^{N}$ with input data $x_i \in \mathbb{R}^p$ and the corresponding binary class labels $d_i \in \{-1, 1\}$, the linear separable optima hyperplane is defined by $(x) = w^T\varphi(x) + b$ where $\varphi(x)$ denotes a fixed-feature space transformation and b a bias parameter. A observation x is assigned to class 1 of $g(x) > 0$ or to -1 if $g(x) < 0$. This is equivalent to $d_i(w^T\varphi(x) + b) \geq 1$, $i = 1, \cdots, N$. Maximizing the margin corresponds to solving

$$\min_{w,b} \frac{1}{2} w^T w \tag{4.4}$$

$$st \quad d_i(w^T\varphi(x) + b) \geq 1, i = 1, \ldots, N \tag{4.5}$$

However, if this formulation is only valid for linear separable classes. For non-linear separable classes, slack variables $\xi_i, i = 1, \ldots, N$ are introduced. These allow penalties to be set for data points wrongly classified. Minimizing the error as

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi \tag{4.6}$$

$$st \quad d_i(w^T\varphi(x) + b) \geq 1 - \xi_i, i = 1, \cdots, N \tag{4.7}$$

$$\xi \geq 0 \tag{4.8}$$

where $C \geq 0$ controls the trade off between the training error and the margin. The dual problem is easier to solve in the feature space. A formulation of the dual problem for a non separable sample of training $\{x_i, y_i\}_{i=1}^{N}$ is presented as

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j k(x_i, x_j) \tag{4.9}$$

$$st \quad \sum_{i=1}^{N} \alpha_i d_i = 0 \tag{4.10}$$

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, N \tag{4.11}$$

where $k(x_i, x_j) = \varphi^T(x_i)\varphi(x_j) = \sum_{i=0}^{m_1} \varphi_i(x_i)\varphi_l(x_j), i = 1, ..., N$ and $j = 1, \ldots, N$. $\varphi_l(x_i)$ corresponds to the l components in the application of $\varphi(x_i)$ from x_i and m_1 defines the dimension of the feature space.

The three most common types of inner-product kernels for SVMs are the polynomial, defined as

$$k(x, x_i) = (x_i \cdot x_j + 1)^d \tag{4.12}$$

the radial-basis function

$$k(x, x_i) = \exp(-\gamma ||x - x_i||^2), \ \gamma \geq 0 \tag{4.13}$$

with γ the parameter that defines how far the influence of a single training example reaches and the hyperbolic defined as

$$k(x, x_i) = \tanh(kx_i \cdot x_j + c) \tag{4.14}$$

with $k > 0$ and $c < 0$.

4.2.2 Naive Bayes

Naive Bayes (NB) (Russell and Norvig, 2016) classifier assumes that features x_j are independent given the class variable C_i. NB is based on Bayes theorem, which provides a mathematical framework for describing the probability of an event being the result of two or more causes. NB is easy to construct, robust and performs quite well, even outperforming more sophisticated alternatives.

$$f_i(X) = \prod_{j=1}^{N} P(x_j|c_i)P(c_i) \tag{4.15}$$

where $X = (x_1, x_2, \ldots, n_N)$ denotes the feature vector, and $C_j, j = 1, 2, \ldots, N$ possible class labels.

4.2.3 Random Forest

Random Forest (RF) (Ho, 1995) are methods that fall in the category of ensemble methods by using multiple decision trees. RF enables to define an order of importance of each attribute into the final model. To train the RF classifier, a set of labeled data is used and by maximizing the Gini index or information gain criteria, the final model is obtained. Its composed by an Ensemble of Decision Trees, most of the time trained with a bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result.

The working principle starts with the generation of a large number of trees that vote for the most popular class. The Strong Law of Large Numbers proves that they always converge so that overfitting is not a problem, and using the Central Limit Theorem, the variance of the sample average has a variance equal to the variance of individual estimator divided by square root of N, attaining low-bias and low-variance properties. Formally, given a collection of tree-structured classifiers $h(x, \theta_k), k = 1, ...$ where the θ_k is independent and identically distributed, random vectors and each tree casts a unit vote for the most popular class for input x (Figure 4.5).

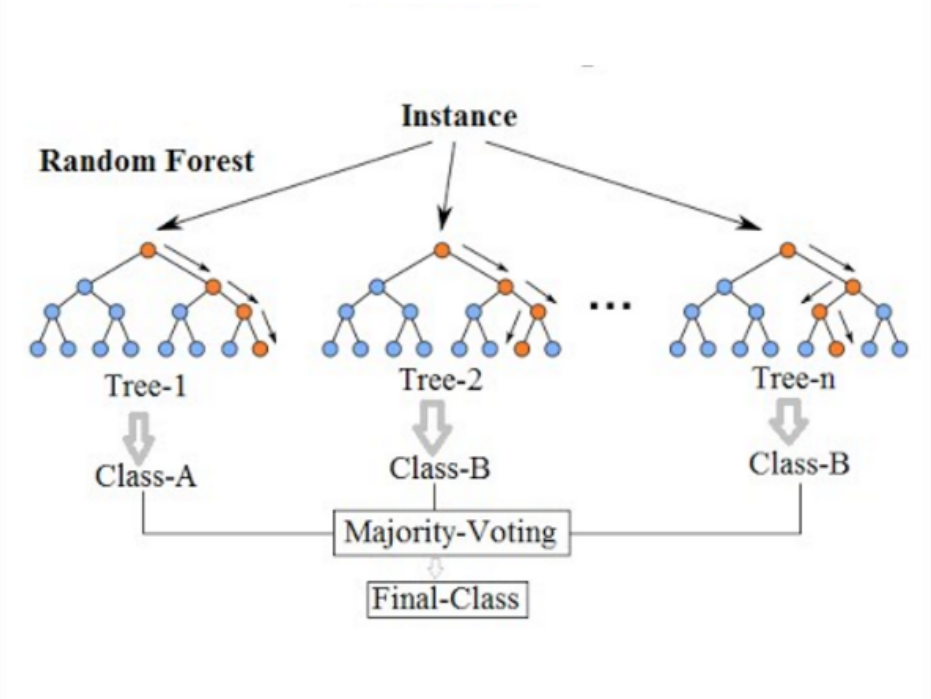


Figure 4.5: Random Forest architecture (Image from [8])

4.2.4 K-Nearest Neighbours

Nearest Neighbours is a non-parametric method with the main objective to estimate the density function from sample patterns. If these estimates are satisfactory, they can be replaced for the true densities when designing the classifier. This particular method extends the local region around a data point x until the k^{th} nearest neighbor is found. This means that for a test point x', the most represented class in the k-closest cases of examples define the predicted class. The design of the classifier lies only in the estimation of the best k, found by using a grid search approach (Hsu et al., 2003) over k to determine which value for this variable gives the lowest error estimation. To determine which point is closer normally the Euclidean Distance (Equation 4.16) is used

$$D(a.b) = \left(\sum_{i=1}^{P} (a_i - b_i)^2 \right)^{1/2} \tag{4.16}$$

where a_i and b_i are the coordinates of all a and b possible points withing dimension P.

[8] https://medium.com/williamkoehrsen/random-forest-simple-explanation-377895a60d2d

4.2.5 Ensemble

Ensembles are commonly composed by multiple weaker models, trained independently with the individual model's predictions being combined at the end. Three types of ensemble models can be found:

1. **Bagging**, also called Bootstrap aggregating, consist in multiple models of the same type with equal weight, trained with random sub-samples of the training dataset. Individual classifiers are trained independently.
2. **Boosting** consisting of multiple models of the same type, with the most recent model learning to fix the predictions errors of the previous model. The training is sequential and iterative, but more prone to over-fitting.
3. **Voting** consist of multiple different models and simple metrics to combine predictions.

A boost classifier can be described in the following

$$H_\tau(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{4.17}$$

where f_k corresponds to the output of a weak learner with input x that corresponds to returned the class of the object. Predicted class is identified by sign and the confidence of classification is given as a absolute value. The total sum training error of the state n of the boost classifier is minimized as

$$E_n = \sum_i E[H_{t-1}(x_i) + \alpha_n h(x_i)]) \tag{4.18}$$

where $H_{t-1}(x)$ is the boosted classifier built on previous steps, $E(H)$ a error function and $f_n(x) = \alpha_n h(x)$ the weak learner to be added to the final classifier. According to to Wu et al. (2008), AdaBoost is one of the most important boosting classifiers, with a solid theoretical foundation, accurate prediction, great simplicity, and successful applications.

4.2.6 Ordinal Classification

Machine learning methods for classification problems as briefly reviewed in the previous sections assume an unordered class distribution corresponding to nominal data problems. However, in many situations as the case of assessing the severity of lesions findings in the BI-RADS scale to assess the final score range comprehended between 0 to 6, the output space exhibits a natural order, an ordinal one.

The standard approach to ordinal data classification encompasses the transformation of the class value to a numeric quantity, and then a regression learner is applied to the transformed data. The numeric output is translated back to the discrete class value. The main disadvantage of this method is that it can only be applied in conjunction with a numeric regression scheme. To overcome this, Frank and Hall (2001) proposed a new method where it firstly transforms the data of a K-class problem to a $K-1$ binary class problem. The training of the i^{th} classifier involves the transformation of the K ordinal class into a binary one where the i^{th} discriminator is obtained by separating the classes $C_1, \ldots, C_i$ and $C_{i+1}, \ldots, C_k$ (Figure 4.6). The i^{th} class represents the test $C_x > C_i$.

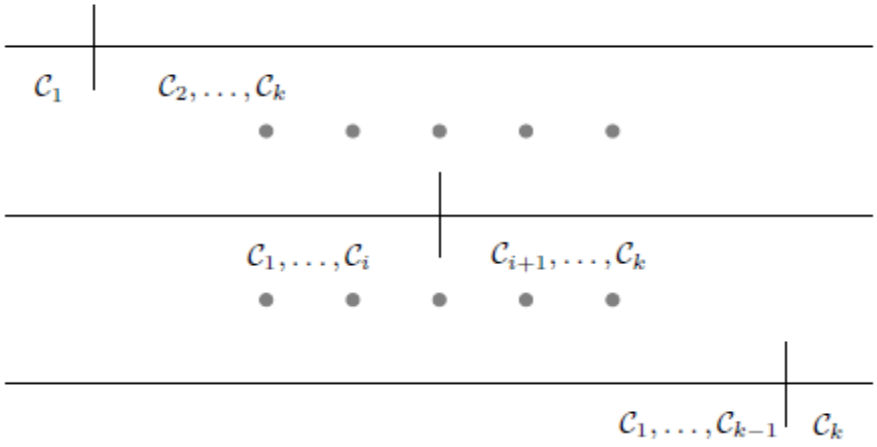


Figure 4.6: Transformation of an ordinal class to binary one.

To predict the class value of an unseen instance, the $K-1$ binary outputs are combined to produce a single estimation. Any binary classifier can be used as the principal method when using this scheme.

4.2.7 SMOTE Resampling

Synthetic Minority Oversampling Technique (SMOTE) for oversampling is a method proposed by Chawla et al. (2002) to address the class in-balance problematic. Often real-world datasets are predominately composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples. Also, the cost of miss-classifying an abnormal (interesting) example as a normal example is often much higher than the cost of the reverse error. SMOTE operates in the feature space to over-sample the minority class by creating synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

Synthetic samples are generated by taking the difference between the feature vector (sam-

ple) under consideration and its nearest neighbor. Next, the difference is by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. The approach effectively forces the decision region of the minority class to become more general.

4.2.8 Experiments and Results for Mass Lesion Classification

For mass classification two main experiments were conducted, binary and multi-class classification. Approaches followed and models comparisons are discussed in detail.

4.2.8.1 Experiments with Mass Lesion Binary Classifier

In order to construct a binary classifier, the threshold for the BI-RADS frontier was set at level 3 using the INbreast database containing masses. According to Table 1.1 the above classes account for the most serious cases, (5 and 6) labeled as malign. The final balanced distribution is obtained (Figure 4.7). The split value was set to 75%, 25% for training for a universe of 143 original cases subject to data replication using scaling $\{0.71.0, 1.3, 1.5\}$, multiplying the training set by 4 while the test set was maintained in its original form and the same subset was evaluated in all experiments.

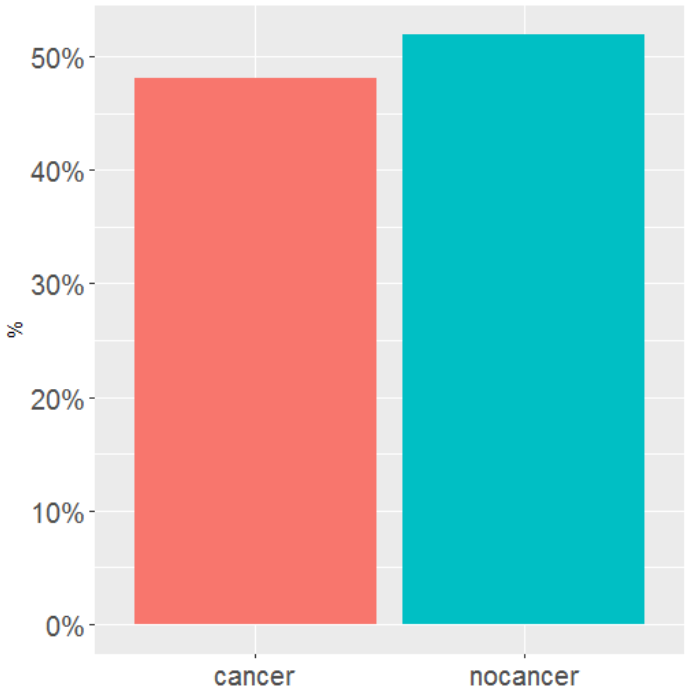


Figure 4.7: Binary Class Distribution (Blue - Benign, Red - Malign).

To access the best model, SVM, Decision Trees, NB and RF classifiers fitted with different

parameters to be fine tuned.

- A group of SVM classifiers with Linear and RBF Kernels, with C ranging from $[1, 5]$ and γ from $[0.1, 2]$ was set.
- Group of RF model was set with $50, 100, 300, 500, 1000$ trees.
- Group of Decision Trees were set with a post-pruning standard error se of $\{0.5, 1\}$ and a $minsplit = 5$

  NB was set with Laplace smoothing

The selected performance metric was the error rate using the Cross Validation (CV) method with 10 k-folds. Figure 4.8 presents the error rate for each of the selected models.

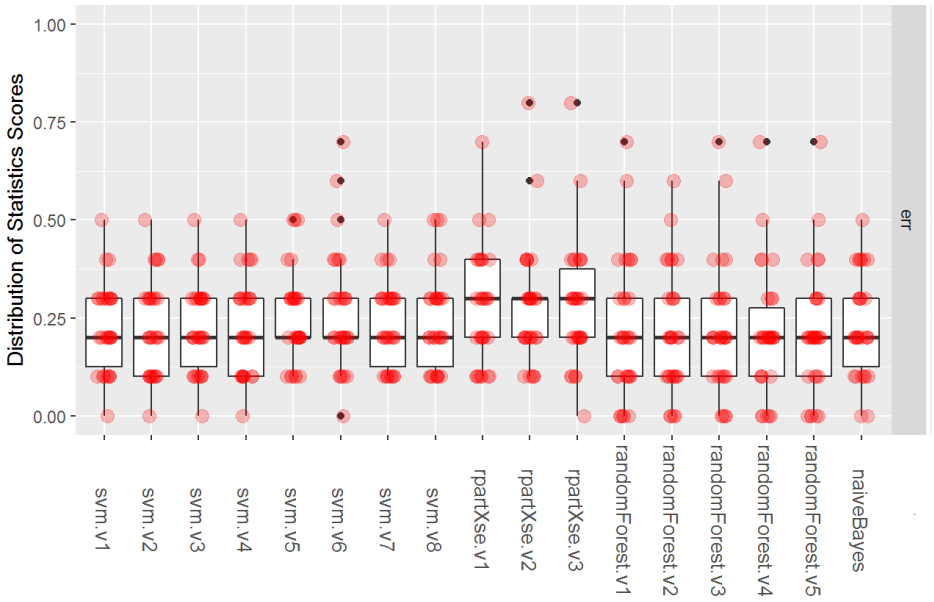
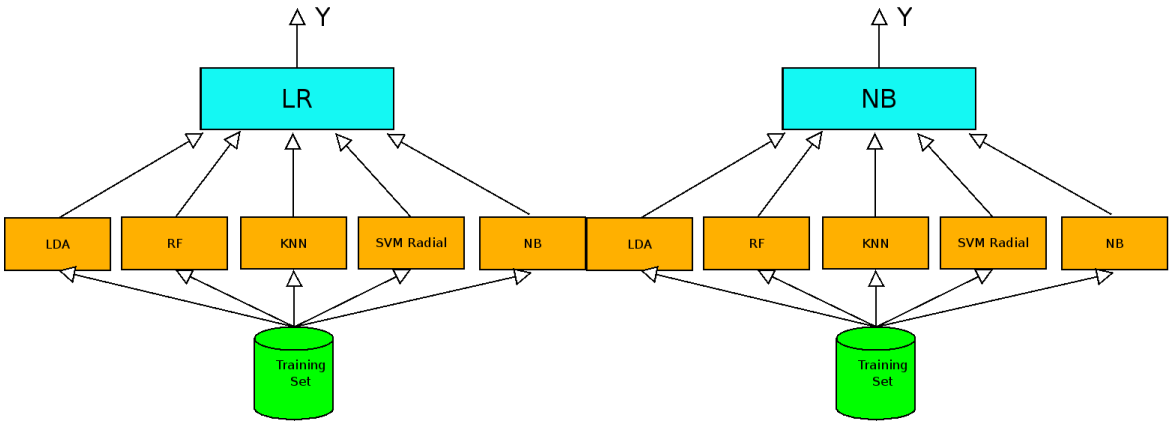


Figure 4.8: Models error rate for binary classification (0 - Benign, 1 - Malign).

The result shows that RF presented a better accuracy when using a number of trees equal to 300, yielding an accuracy of 0.801.

In order to increase the robustness of the classifier stage a meta-models formed by an Logistic Regression (LR) and a NB to ensemble collections of predictive models (SVM, Knn, NB, Linear Descriminant (LDA) and RF are defined and optimized (Figure 4.9). Model evaluation was CV with 10 k folds using accuracy metrics.



Ensemble using LR Ensemble using NB

Figure 4.9: Ensembles stacking architecture.

Figure 4.10 presents accuracy and Kappa statistics for individual models. RF with 300 trees holds the best Kappa.

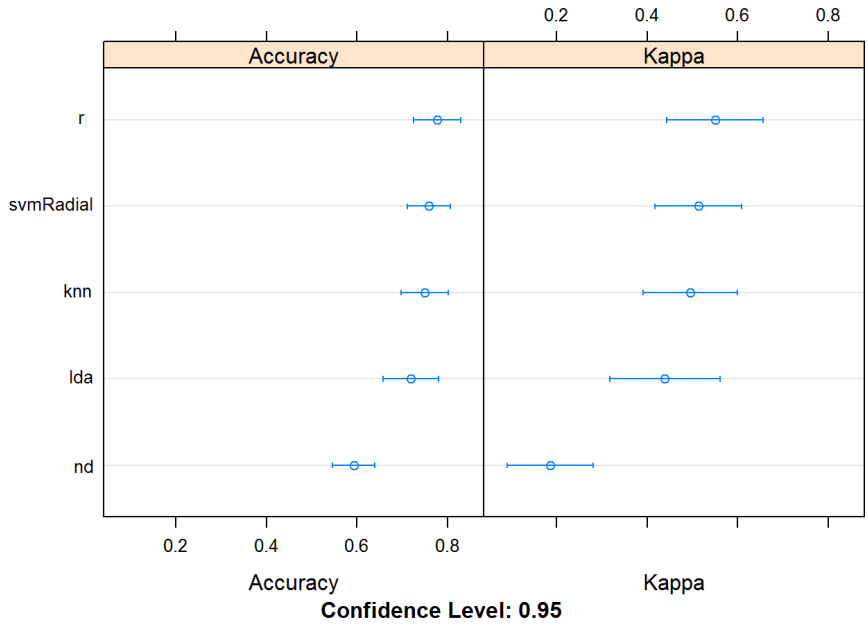


Figure 4.10: Individual accuracy for each of models in the ensemble (confidence level 0.95).

Table 4.3 presents accuracy comparison between the two staking methods, with the same base models and parameters.

Table 4.3: Comparison between two ensemble models.

Stacker	Acc	Kappa
Single RF	0.801	-
NB	0.824	0.521
LR	0.835	0.673

Acc is a measure of accuracy (higher the better), Measures range $[0, 1]$

The ensemble with an LR stacker presents better results when compared with NB and yielded almost a 3.5% accuracy increase when compared with the best RF single model. NB on the other yielded almost the same performance of the original model, showing that this stacker choice was not the best one. Ensemble models with LR stacker were able to capture data variance that was missed by RF single models.

4.2.8.2 BI-RADS Model Evaluation

For models be more discriminatory its important that the BI-RADS level can be accurately predicted, providing clinicians with more insights about lesion manifestations and un-hide potential positive cases that can be missed when using simple binary classification. For the task, models were trained using all the available BI-RADS levels. Analysis of class distribution (Figure 4.11) exhibits strong in-balance in higher classes that may lead to poorer discriminative models. The split of the dataset was also 75%-25% for training and testing using the INbreast dataset with a universe of 116 mass cases. Same data replication and test evaluation using in the binary experiment was set in place.

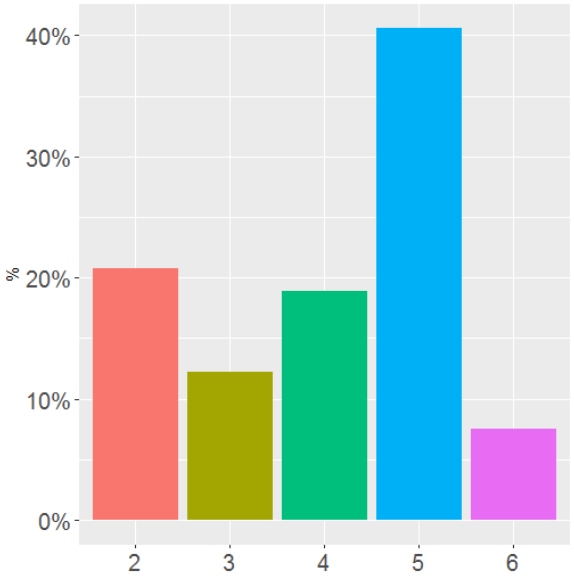


Figure 4.11: BI-RADS class distribution.

To set a baseline model, a simple SVM classifier with RBF kernel with with parameters ranging from $C = \{1, 10\}$ and $gamma = \{0.01, 0.5\}$ is fined tuned, yielding Mean Absolute Error (MAE) of 1.921 with parameters $C = 2$, $gamma = 0.02$.The confusion matrix is presented in Figure 4.12.

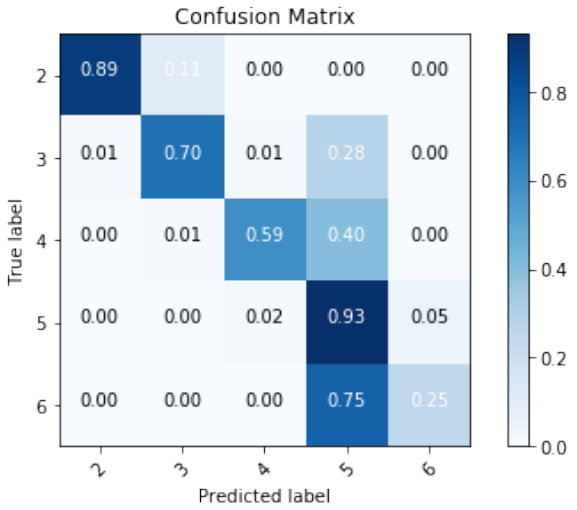


Figure 4.12: Confusion Matrix with 5 Classes.

Results show that 28% of the BI-RADS 3 were classified as 5, 40% of the BI-RADS 4 were classified as 5 and 75% of the BI-RADS 6 were classified as 5. BI-RADS 5 and 6 exhibit a large percentage of miss-classification with the same describing features being strong correlated.

To avoid miss-classification between the BI-RADS 5 and 6, both classes were merged into a single one and a new experiment was conducted. This decision was made since BI-RADS 5 corresponds to highly suggestive of malignancy where a Biopsy is required and BI-RADS 6 corresponds to proven malignancy after biopsy corresponding almost to the same malignancy category. The new class distribution is represented on Figure 4.13. In addition, for model and parameter search in the Grid Search, the class weight was set as an input argument to be the inverse class weight related to the train class frequency.

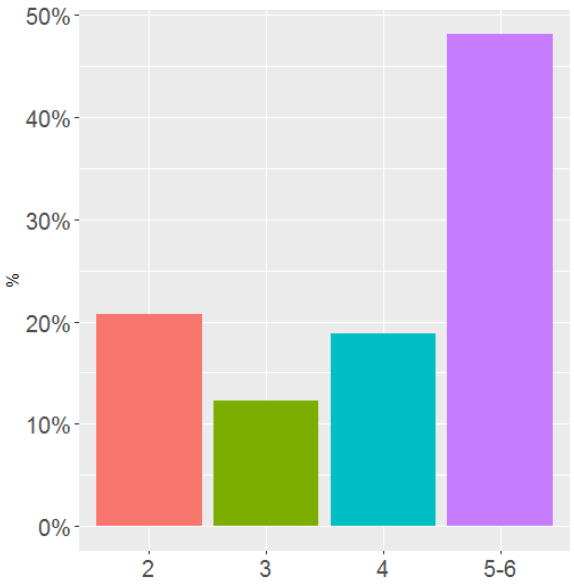


Figure 4.13: BI-RADS class distribution (BI-RADS 5 and 6 merged).

To access the impact of re-sampling data with the newly merged BI-RADS class, a set of models RF, Extreme Randomized Trees (ERT), SVM and k-Nearest Neighbours (kNN)) where trained using the original data and re-sampled data by employing Synthetic Minority Over-sampling Technique (SMOTE) method. Evaluation method was CV with 10 k-folds to maximize accuracy, with models parameters set as to be:

- A set of SVM models using Linear and RBF kernels, with C ranging from $[1, 5]$ and γ of $\{0.01, 0.5\}$
- A set of RF models with number of trees of $\{50, 100, 300, 500, 1000\}$ and number of splits $\{2, 5\}$.
- A set of kNN models with k set to $\{3, 5, 9\}$ (odd number to avoid ties)
- A set of ERT models with a number trees of $\{50, 100, 300, 500, 1000\}$ with a number of splits $\{2, 5\}$

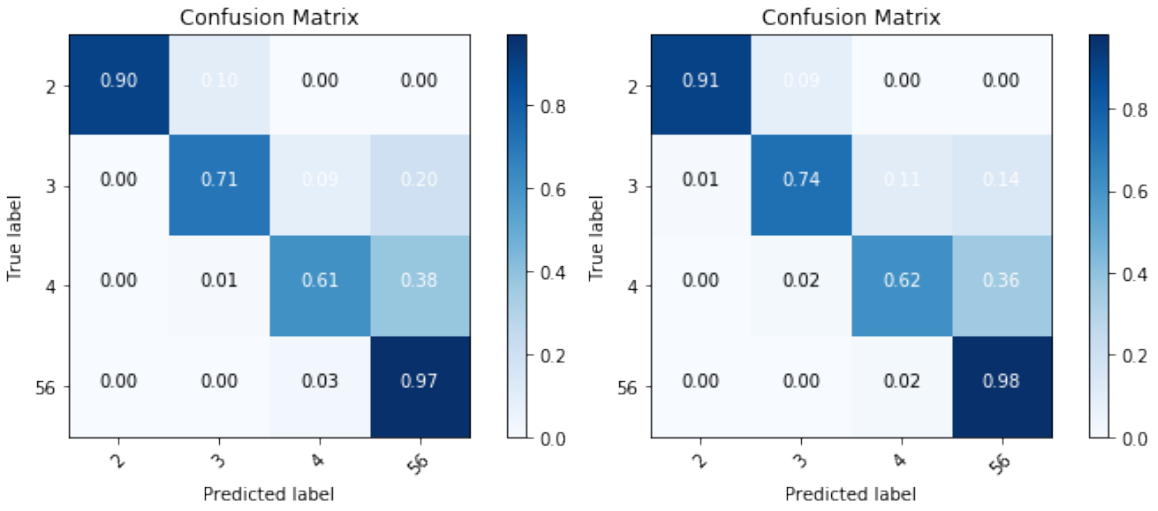
Performance metrics for both cases with best models are summarized in Table 4.4. No weight matrix was considered when computing the accuracy.

Using re-sampled data with the merged class, the best model was RF with a number of trees of 300 with a split value of 2, while in using the original data, an ERT model with a number of trees of 500 and a split number of 2 presented the best result. The model trained in re-sampled data presented a lower MAE. Figure 4.14 presents a pairwise confusion matrix between the two best performer models.

Table 4.4: Comparison between non and pre-processed data.

Data	Model	MAE	Acc
Baseline	SVM	1.912	0.672
Original 56	ERT	1.010	0.798
SMOTE	RF	0.891	0.813

MAE is a measure of error (lower the better), Acc is a measures of accuracy (higher the better), Measures range $[0, 1]$.



Without SMOTE With SMOTE

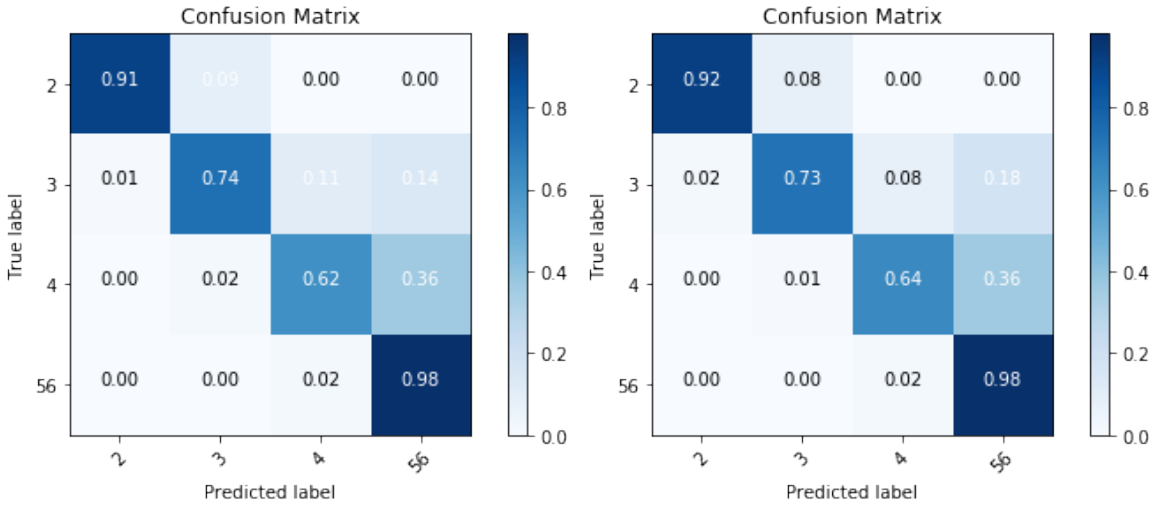
Figure 4.14: Confusion Matrix for BI-RADS class classification with non re-sampled data vs SMOTE re-sampled data.

Careful analysis allows to conclude that inter-class miss-classification appear frequently in adjacent BI-RADS classes. Also the merged of the upper classes lead to a significant improvement in the initial models, improved by the new re-sampled data. Majority of the machine learning methods are biased towards the major class, showing a great sensibility to class in-balance. ERT presented better result when compared to RF on non resampled data since ERT have was fitted using bootstrap re-sampling.

The several comparisons on the models show an effective increase of performance when using re-sampling techniques and merging BI-RADS 5 and 6 classes.

To access the influence of data ordinality, an experiment was conducted using the SMOTE re-sampled data fitted by a using Ordinal Regression Trees (ORT). The advantage when applying ordinal regression trees is that the power of the statistical test to correctly detect an association between a predictor and the ordinal response is higher. It is thus less likely that a noise predictor yields a lower p-value just by chance and is selected for the split. The selected prediction performance to perform the pruning of the classification tree was the

total miss-classification cost. A model was trained using the same resampled set and same evaluation methods, yielding a final MAE of 0.661. Figure 4.15 presents the final confusion matrix.



With SMOTE | With SMOTE + Ordinal

Figure 4.15: Confusion Matrix for BI-RADS class classification with SMOTE re-sampled data vs ordinal classifier.

The confusion matrix exhibits an improvement in the inter-class error namely in upper BI-RADS classes. This small improvement is due to the different criteria that try to minimize large distance classes error. No weight matrix was considered when computing the accuracy. Table 4.5 summarizes all experiments.

Table 4.5: Overall Comparison

Data	Model	MAE	Acc
Baseline	SVM	1.912	0.672
Original 56	ERT	1.010	0.798
SMOTE	RF	0.891	0.813
SMOTE + Ordinal	ORT	0.842	0.820

MAE is a measure of error (lower the better), Acc is a measures of accuracy (higher the better), Measures range $[0, 1]$.

A plausible explanation for the improvement in the ranking by using ordinal regression trees is that in ordinal regression trees it is more likely that a predictor associated with the response is selected for a split. A predictor that is often selected in a tree and occurs close to the root node of the tree is likely to receive a high importance. These results enable to conclude that accounting for class in-balance and ordinality enables to obtain more accurate models. Also, the BI-RADS 3 and 4 show always some higher predicted value than its true value, this is due to the presence of masses and calcification's on the same image, with the

report describing only the higher BI-RADS class that corresponds to the calcification. Two different models considering only masses and calcifications should be employed to determine the highest BI-RADS from both genres of lesions.

4.3 Deep Learning for Breast Classification

Deep learning is part of a broader family of machine learning methods that are based on learning data representation, as opposed to task-specific algorithms. Deep learning models are vaguely inspired by information processing and communication patterns present in biological nervous systems, however with various differences from the structural and functional properties of the biological brains, making deep learning incompatible with neuroscience evidence. Deep learning can be described as a class of machine learning algorithms that:

- use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the outputs from the previous layer as its input,
- can be supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis),
- learn multiple levels of representations that correspond to different levels of abstraction.

4.3.1 Feed Forward Artificial Neural Networks

Artificial Neural Networks (ANN) is composed of simple units (Neurons) responsible for computing an output (activation) from its inputs that correspond to outputs from previous neurons. The more common ANN is the fully connected feed forward neural network, used to solve problems of classification or regression by approximating the network output to the real class/value, for each input data. The feedforward network corresponds to sequential layers, where a unit of layer k receives as inputs all neurons of layer $k-1$ (Figure 4.16).

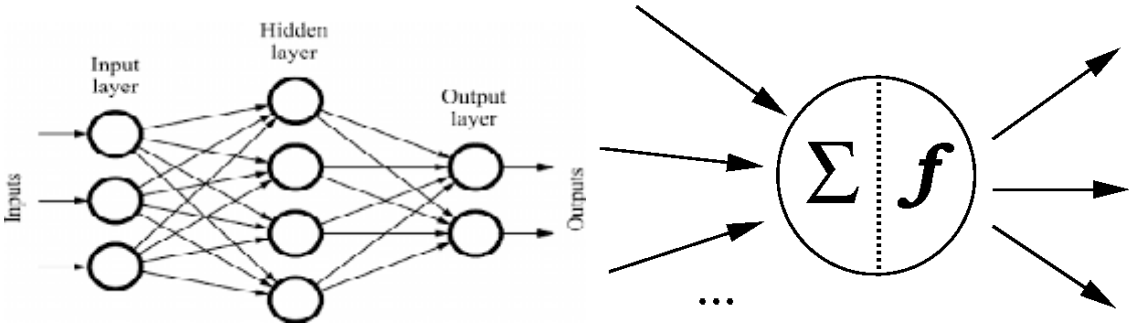


Figure 4.16: Feed forward ANN and neuron unit.

The basic computational element (model neuron) is often called a node or unit. It receives input from some other units, or external source. Each input has a associated weight $w_{k,i}$, that can be modified, modeling a synaptic learning. The unit computes a activation function f that uses the weighted sum of its inputs $y_i = f(\sum_i w_{k,i}^T * I_{k-1} + b_{k,i})$, being f a non-linear activation function that can take may forms, and $w_{i,j}$ and $b_{k,i}$ are weights and bias learned during training.

4.3.2 Convolutional Neural Networks

CNN are models composed mainly of two parts, the convolutional and fully connected parts. The first acts in the spatial image space to extract features while the second corresponds to a fully connected feed forward ANN. In the convolutional part, each layer is organized in a 3D volume *(width, height and depth)* (Figure 4.17).

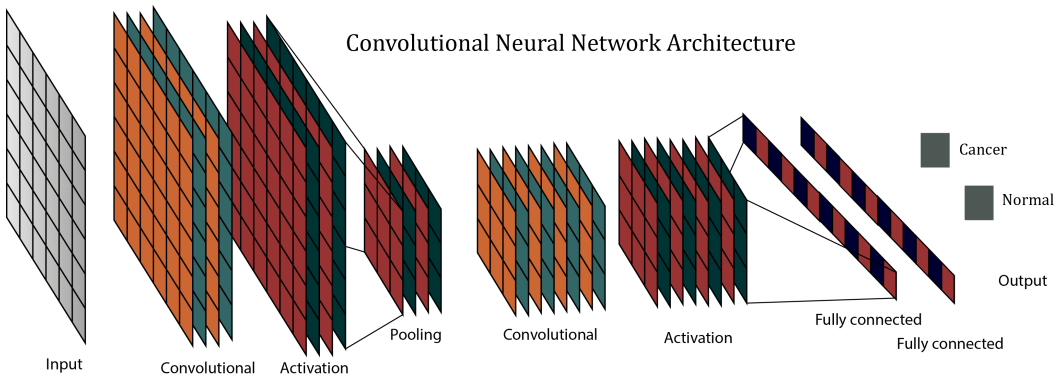


Figure 4.17: Diagram of a CNN Architecture for Benign - Malign classification.

Each neuron receives inputs from the previous layer inside a specific interval of *width* and *height*, defined as the receptive field. All neurons in the same layer share the same

parameters and the output neuron are independent of the vertical and horizontal positions, making then invariant to shifts. A series of spatial filters are applied to the previous layer, with the number of filters being equal to the depth of the current layer. Filtering level on the image from each layer is controlled a stride parameter s that controls the step size, resulting in a height or width of $1/s$ in the respective layer. On CNNs, later layers are trained on feature representation computed in earlier ones, enabling the CNN to learn a hierarchy of signals components. Low level extracted components are used to obtain higher level components and so on. The final fully connected part is just a simple portion of the network that does not have the same spatial and parameter restriction as the convolutional part.

4.3.2.1 Layers

The majority of CNN layers have parameters and spatial restrictions. However, the transformation they apply to the input may differ. Each of the layers are presented in detail on the following subsection.

4.3.2.2 Input Layer

The input layer is responsible for feeding the input data into the model. The main restriction relies on the fact that input data must have a fixed input shape. RGB or gray-scale is the common image data representation, with the data commonly being converted into a 3 or 1 dimensional array according to the cases, resulting in shapes $[w, h, 1]$ or $[w, h, 1]$, where w corresponds to the width of the image, h to height and the last column the number of color channels.

4.3.2.3 Convolutional Layer

Considering the same example, neuron in position $[x, y, z]$ is connected to $3 \times 3 \times f_{-1}$ inputs, with f_{-1} symbolizing the number of filters from the previous layer. The activation of this particular unit is obtained by multiplying the output of each neuron of the receptive field by the corresponding weight connection and summing all values with a bias term. The restriction parameter on CNN is related to these weights. Neurons in the same layer and same depths share the same weights and bias. However, due to the fact that they are located in different (x, y) positions, they belong to different receptive fields, resulting in different activation's. Each neuron performs a linear combination of its receptive field and this operation is shifted throughout the whole size (w, h) of the corresponding layer. For each value of z

$$R_z = \sum_{c=1}^{f_-} (L_c \circledast_s F_{z,c}) + b_z, \text{with} \quad R = (R_1, R_2 \cdot \cdot, R_f), \tag{4.19}$$

where $(\circledast_s)$ is the strided convolution, R is the concatenation of the partial results of propagating the input feature map L with L_- filters, in a convolutional layer with stride s containing f filters, F weights and bias b. Each convolutional layer with a filter of a size of m shrinks the output, relative to the input by $(m-1)$ pixels. To account for this reduction, the image is padded with $(m-1)/2$ zeros on each border. In the majority of the architectures, f grows as spatial resolution decreases. This is due to the fact that earlier convolutional layers have more general features, and deeper convolutional layers have smaller resolutions, containing more specific high-level features that require a higher dimensional space. In addition, earlier feature maps stages have bigger resolutions, requiring more memory, thus penalizing the use of higher values of f. The number of weights to optimize is directly related to the size of the employed filter m. m is kept normally small, $(3, 3)$ that can be staked by another equal filter, resulting in an effective filter of size $(5, 5)$. This configuration has the advantage of reducing the number of weights to optimize $(2 \times (3 \times 3 \times f))$ against $(5 \times 5 \times f)$. It also enables to introduce non-linear functions between both filters, resulting in a more discriminative function (Simonyan and Zisserman, 2014b). As a disadvantage, this procedure requires larger portions of memory to maintain the internal representations of the intermediate feature maps. One is the more common value choice for stride s. Higher strides are manly used on first convolutional layers to reduce the spatial resolution quickly and save memory, however, inappropriate value choice can lead to loss of relevant image information.

4.3.2.4 Activation Function

The main objective of the Activation Functions is to introduce nonlinearities in the model, making then more discriminative by avoiding that the outputs become a simple linear combination of the inputs. They perform a simple element-wise operation in the model, conserving respective layer size. To facilitate training, the Rectified Linear Unit (ReLU) has been employed as the common standard on CNN in detriment of Sigmoid functions, resulting in faster learning times. ReLU can be defined as

$$\text{ReLU} \longrightarrow r_{m,n,c} = \max(0, I_{x,y,z}) \tag{4.20}$$

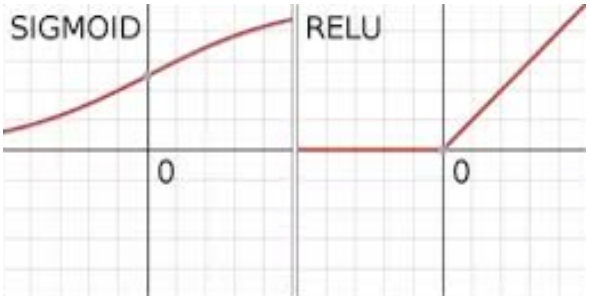


Figure 4.18: Two common activation's functions used in CNNs.

CNN may exhibit dying out phenomena since ReLU derivative for input $r_{k,i,j} < 0$ is equal to 0, leads to the output never being activated. To circumvent the problematic, careful initialization of the weights and adequate leaning rates can eliminate this phenomena, or alternatively, a leaky ReLU can be used. A Leaky ReLU returns $\max(\alpha \times I_{k,i,j}, r_{k,i,j}$ with α being defined to lower value, avoiding outputs equal to zero.

4.3.2.5 Polling Layer

Pooling enables to further modify the output of the layer. A pooling function replaces the output of a convolution layer at a certain location with a summary statistic of the nearby outputs (Figure 4.19).

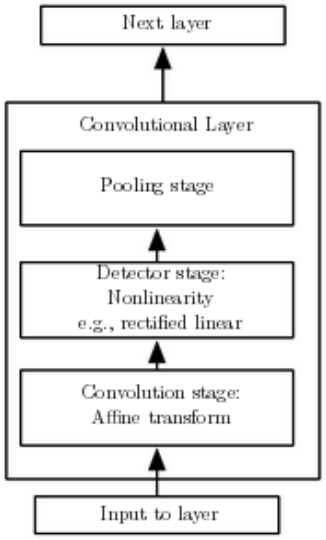


Figure 4.19: Max-pool operation on a convolution layer stage.

Max-Pooling (Zhou et al., 1988) is commonly used in CNN. Max-polling is a non-linear op-

eration that reports the maximum output within a rectangular neighborhood (Figure 4.20). A max-polling operation can be formulated as

$$r_{x,y,z} = \max(I_i, j, z), \text{with} i \in [s \times x, s \times x + m[, j \in [s \times y, s \times y + m[\tag{4.21}$$

Other popular pooling functions can include the average of a rectangular neighborhood, L^2 norm of a rectangular neighborhood or a weighted average based on the distance from the central pixel. The main purpose is to reduce the spatial size of the network while providing some invariance to translation. With this operation each neuron outputs the maximum for a small region of the input, according to the filter size, adding an extra control parameter to avoid over-fitting.

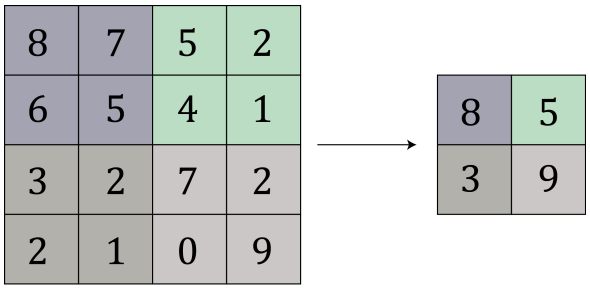


Figure 4.20: Max-pool operation on a small 2-dimensional array. In this case, m = 2 and s = 2.

The spatial reduction allows the next layer units to be influenced by larger regions of the original input image instead of small and not representative regions.

4.3.2.6 Dense Layer

Dense layers work similarly as the hidden layer on a fully-connected feed-forward neural network. If a previous layers output has a spatial structure, a flattening operation is performed to reshape the rank 3D tensor with size $[w, h, d]$ into a rank 1D with size$[w \times h \times d]$. The number of weights is given by $n_l \times (n_{l-1} + 1$. where n_l corresponds to the number of units of the layer l. Bias corresponds to an additional weight per neuron. In detail, the dense layer performs the following operation

$$\begin{aligned} o_i &= w_i^T . I_{k-1+b_{k,i}} \\ o &= (o_1, o_2, \cdots, o_d) \end{aligned} \tag{4.22}$$

resulting in a vector with d linear combinations of all inputs. Considering a convolutional layer with the same filter size as input, dense filters will only be applied in one specific position, namely, the one were they completely fit the input feature map. Parameters are not shared among different depths and no weight is reused. If required, dense layers can be transformed into convolution ones just by reshaping the weights of the input, enabling to transform a classification model into a screening model containing only by convolutional layers.

4.3.2.7 Output Layer

The output layer is normally a linear combination of the inputs, coupled with a non-linear function between the last fully-connected layer and the output neurons. For classification tasks, the output of the model corresponds to a set of values, each one representing the probability of the input belonging to a specific class $k \in K$. Two main output activation functions are commonly used, Sigmoid and Softmax. Softmax function are commonly employed for multi-class classification by taking an N-dimensional vector of real numbers and transforming into a real vector number that range between $(0, 1)$, where the sum of all classes adds up to 1, $p_j = \frac{e^{a_i}}{\sum_{k=1}^{N} e_k^a}$. Sigmoid activation functions are suitable for binary classification with the probability for the class j given as $p_j = \frac{1}{1+(e^(-x))}$.

4.3.2.8 Dropout

Dropout is a common technique for regularizing ANN, including deep learning models. Proposed by Srivastava et al. (2014), dropout enables to build more robust features by preventing neurons from co-adapting. Dropout is usually staked after the activation functions. The regularization is performed by randomly setting some entries of the input feature map to zero, increasing the difficulty during training. Formally, each of these features has an independent probability σ of being kept, being re-scaled by $1 = \sigma$. Discarded points are set to zero. σ usually ranges from $[0, 1]$ for the training set and 1 for the test set. This enables that on each interaction, some of the neurons from the network being removed temporally along with its input and output layers, forming a slightly different network to be trained. This resembles the concept of RF, were averaging combinations of different models increases the performance of the final model. Dropping penalizes neurons that rely on fewer input connections, favoring larger connected neurons and allowing more general features to be kept.

4.3.2.9 Batch Normalization

Batch Normalization proved to be a very effective method to reduce training times by addressing the phenomenon of internal co-variate shift identified by Ioffe and Szegedy (2015). Co-variate shift slows down training, requiring careful weight initialization. To address the problem, on each training batch, data is normalized with zero mean and variance 1 (Equation 4.23 for all neurons in the same depth. The mean and Standard Deviation (STD) are usually referred to mini-batch statistics. In addition, a running average of these values is kept to be used during inference, avoiding that the output for a new example becomes dependable on the mini-batch statistic and affected by other inputs running in parallel. To ensure that the model, in a particular stage is able to represent the same function with or without batch normalization, new trainable weights can be added, γ and β to scale and offset the output respectively. These new weights can be defined as

$$\begin{aligned} I_c &= \gamma \times \frac{I_c - mean(I_c)}{std(I_c}, \quad \text{Training} \\ I_c &= \gamma \times \frac{i_c - u_c}{v_c} + \beta, \quad \text{Inference} \end{aligned} \tag{4.23}$$

where u_c and v_c are the running averages for the $mean(I_c)$ and $std(I_c)$. Batch normalization have allowed the use of higher learning rates, leading to the reduction the required number of interactions for model convergence. Batch normalization layer can be employed between linear and activation layers, however Mishkin et al. (2016) suggest the use of batch normalization after activation layers to improve model accuracy.

4.3.3 Optimization

Considering a binary classification task using a dataset D containing N images $I \in \mathbb{R}$, with each of the images containing a associated label $y \in \{0, 1\}$. Given image I_i, a model must predict a label y with an associated probability $p(I_i)$. To optimize neural networks many alternatives can be employed. The most commonly used technique rely in the minimization of a loss function by means of gradient descent. In multi-class classification, cross entropy is used as loss functions. Cross-entropy measures how well one distribution probability approximates another for a given set of events, in this case, I images. To access how well $p(I_i)$ approximates the real distribution of label y_i, the loss function L defined as

$$L = -\frac{1}{|D|} \sum_{i}^{|D|} (y_i \log(p_(I_i)) + (1 - y_i) \log(1 - p(I_i))) \tag{4.24}$$

is evaluated on each interaction to be minimized. The probability of a input depends only on its weights θ and can be defined as $p(I, \theta)$. Given θ, is possible to compute $L(\theta)$ by running the model on the dataset D and extract the cross entropy.

4.3.3.1 Back-propagation

In order to compute the network error and adjust weights, computation of the gradient of the loss function regarding the weights $\nabla_\theta L(\theta)$ is performed during training. This process is called back-propagation, were first, the input data is propagated among the networks and the loss $L(\theta)$ is computed in the forward pass, and second, the loss is propagated through all the weights of the network. The gradient with respect to the output is given by

$$\begin{aligned} \frac{\partial L}{\partial p} &= \frac{\partial(-(y\log(p) + (1-y)\log(1-p)))}{\partial p} \\ -\frac{y}{p} - \frac{1-y}{1-p} &= \frac{-(1-p)y + p(1-y)}{p(1-p} = \frac{p-y}{p(1-p} \end{aligned} \tag{4.25}$$

Considering the derivative of the *sigmoid* function with respect to its inputs, i as

$$\frac{\partial sigm(i)}{\partial i} = sigm(i)(1 - sigm(i)) \tag{4.26}$$

, and using the chain rule for derivatives, is possible to obtain the derivative of L with respect to i as

$$\frac{\partial L}{\partial i} = \frac{\partial L}{\partial p}\frac{\partial p}{\partial i} = (p - y). \tag{4.27}$$

The same principle can be used to obtain the gradients of the loss function with respect to the weights for the last fully connected layer. Considering the fact that this layer computes a linear combination of its inputs, a, the derivative of i with respect to the weights, w becomes a vector a as

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial p}\frac{\partial p}{\partial i}\frac{\partial i}{\partial w} = (p - y)a. \tag{4.28}$$

enabling to compute the $\nabla_\theta L(\theta)$ in a efficient manner.

4.3.3.2 Gradient Descent

Gradient descent starts by computing the derivative $\frac{\partial f}{\partial x}$ of a differentiable function $f(x)$ to determine the x value that makes $f(x)$ minimum. The derivative informs in which direction the function is increasing, enabling to interactively update x so the derivative $\frac{\partial f}{\partial x}$ converges towards the minimum. The same approach is performed on multi-variable derivatives, where the main objective is to minimize a loss function $L(\theta)$ by interactively update θ towards the opposite direction of the gradient. Initially, θ_0 is set to random values to avoid zero initialization. On each interaction, the gradient of L is computed with respect to the current weights $\nabla_\theta L(\theta)$, and on the next iteration the weights are set to $\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta)$, with η as the learning rate. Appropriate η value must be properly set to avoid slow learning process when the value is low, or too high, resulting in large updates that lead to non-convergence. The gradient descent on its original formulation does not grants that the absolute best solution to be found. However, LeCun et al. (2015) states that the majority of the cases, a good local minimum may be satisfactory.

In deep learning, the gradient is not computed for the whole dataset, but rather to a small portion defined as the batch. A batch is extracted from dataset D on each iteration and is used as an approximation of the gradient for the whole dataset D. Bigger batches lead to better estimation of the gradient of the complete dataset, requiring the use of larger learning rates (Mishkin et al., 2016). This genre of batch approach is defined as mini-batch gradient descent. When the batch size is one results in the stochastic gradient descent.

To reduce the training time and allow careful optimization of the weights θ when close to the local minima, a dynamic learning rate η is frequently used. The most common strategy for modulating the learning rate is the use of an exponentially decaying learning rate that reduces the range of the steps near the final objective refining final convergence.

4.3.3.3 Adam

Adaptive Moment Estimation, or ADAM, is derived from Gradient descent with some additions for faster convergence. It incorporates of a momentum term into the update equation

$$\begin{aligned} v_t &= \gamma \; v_{t-1} + \eta \nabla_\theta L(\theta) \\ \theta_{t+1} &= \theta - v_t \end{aligned} \tag{4.29}$$

, enabling a weight $theta_i$ to maintain the same direction in successive updates, increasing the steps towards that direction. If no momentum exists, the weights that have been oscillating

tend to be changed more slowly. The concept can resemble a "ball" loss $L(\theta)$ over a mountain valley, where each point is a set of particular values for *theta* and the height is the current value for the loss function $L(\theta)$. With a simple gradient, when entering a valley, the rate of descent will be constant for many iterations, while the addition of momentum enables to gain speed towards the minimum of the valley. Similar to momentum, a running average of the past gradients are kept. Adam incorporates the principles present in algorithms like Adadelta (Zeiler, 2012) and RMSprop (Tieleman and Hinton, 2012), that favors the update of weights that have not been frequently updated. Formally, for one parameter θ_i and considering the gradient at a particular time t, $g_{i,t} = \nabla_\theta L(\theta_i)$ as

$$\begin{aligned} m_{i,t} &= \beta_1\ m_{i,t-1} + (1-\beta_1) g_{i,t} \\ v_{i,t} &= \beta_2\ v_{i,t-1} + (1-\beta_2) g_{i,t}^2 \end{aligned} \tag{4.30}$$

where $m_{i,t}$ and $v_{i,t}$ are the estimated values for the gradients and squared gradients respectively. β_1, β_2 are tuning parameters. At time $t = 1$, $m_{i,t=1}$ and $v_{i,t=1}$ are zero. To circumvent the initial zero condition, the following correction is done as

$$\begin{aligned} \bar{m_{i,t}} &= \frac{m_{i,t}}{1-\beta_1} \\ \bar{v_{i,t}} &= \frac{v_{i,t}}{1-\beta_2} \end{aligned} \tag{4.31}$$

and the update equation for each weight becomes

$$\theta_{i,t+1} = \theta_{i_t} - \frac{\eta}{\sqrt{\bar{v_{i,t}}} + \epsilon} \bar{m_{i,t}} \tag{4.32}$$

4.3.3.4 Regularization

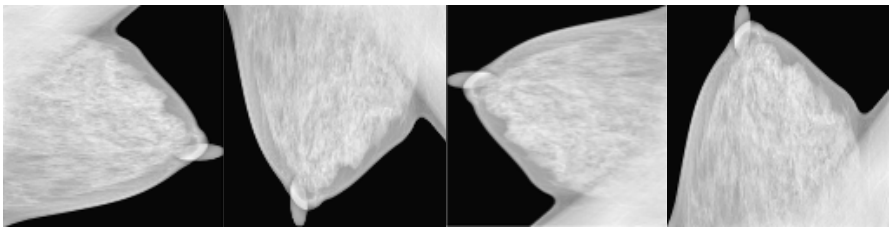
To improve generalization of a CNN model is frequent to use regularization. The more commonly used is the L^2 regularization, enabling high weights to become penalized in the loss function, forcing the model to rely on the majority of the features instead of a small subset by simply adding the regularization term λ to the loss function

$$L'(\theta) = L(\theta) + \lambda\theta^2. \tag{4.33}$$

The λ or weight decay constant term defines how aggressive the L^2 regularization must be, favouring simpler models. In many situations, the term 2 is neglected since the derivative of Equation 4.33 becomes simply $2\lambda\theta$. The introduced λ term makes the weights to exponentially decay on each interaction creating models with better generalization capability and better training accuracy (Krizhevsky et al., 2012).

4.3.4 Dataset Augmentation

The best way to make a machine learning model generalize better is to train it on more data. In the specific cases, the number of samples can be limited. One way to get around this problem is to create synthetic data and add it to the training set. In a classification task, a classifier needs to take a complicated, high-dimensional input x and summarize it with a single category identity y. This means that the main task facing a classifier is to be invariant to a wide variety of transformations. New (x, y) image pairs can be obtained easily just by transforming the x inputs of the training set (Figure 4.21). Operations like translating the training images a few pixels in each direction can often greatly improve generalization, even if the model has already been designed to be partially translation invariant by using the convolution and pooling techniques (4.3.2). Many other operations, such as rotating the image or scaling the image can be also effective.



Original 90° 180° 270°

Figure 4.21: Example of the random augmented images.

4.3.5 Experiments and Results for Deep Leaning Methods For Segmentation

In deep leaning methods, two main tasks were conducted. First the construction of a binary patch classifier model with images extracted from the original dataset. Second the reuse of the pre-trained model with slight modifications to obtain heatmaps (region proposal) of potential mass lesion regions to be classified by the patch classifier model and a final BI-RADS classifier by reusing a pre-trained CNN network.

4.3.5.1 Dataset Construction

Full images were resized to 1/4 of the size, enabling to the majority of the mass lesions to fit totally inside a 112×112 bounding box. For each mammogram images only breast box region was kept and intensity was normalized on range $[0, 1]$ on each image

For constructing the patch dataset, the approach was to extract 40 samples patches from the mass region (lesion) using a bounding box with a 20% of area increase to accommodate for neighbor surroundings and 40 from background area from images containing masses from INbreast database. For mass patches, in particular, an overlapping of 0.9 was used enabling to obtain images with slight differences. This approach enabled to obtain a binary labeled dataset with 44,800 patch images without any augmentation. Figure 4.22 shows some of the samples obtained from the breast images employed in model training.

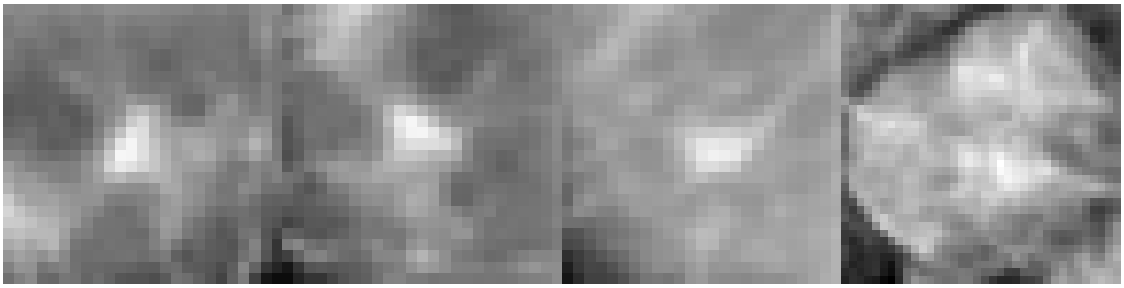


Figure 4.22: Example of sampled patches for masses.

4.3.5.2 Patch Image Class CNN Evaluation

For the patch classification, the main objective is to learn the differences between background and mass lesions. The constructed dataset is divided into train and validation with a split value of 75% and 10% respectively. The test set is extracted from the set with a split of 15%. The employed architectures consist of the Visual Geometry Group (VGG) schemes (Simonyan and Zisserman, 2014b). VGG architecture is composed by an evaluation of convolutional layers with increasing depth using very small (3×3) convolution filters. Three variations of the VGG architecture were implemented end evaluated.

The dataset was constructed from INbreast database to create two distinguishable classes (mass/ non-mass) labeled with 1 and -1 respectively. The output layer is a *tanh* that corresponds to rescaling of the logistic *sigmoid*, such that its outputs range from $[-1, 1]$. Initial weights for the incoming connections to a unit are drawn from a normal distribution with zero mean and $\sqrt{2/n_{im}}$ standard deviation where n_{im} is the number of connections. Biases are initialized to zero. The following Tables (4.6) resumes the main layers used on each of the CNN.

Table 4.6: Description of the first model architecture used for the patch classifier. All Convolutional and Dense layers are followed by a ReLU activation. The output layer has a Hyperbolic Tangent (Tanh) activation function for binary classification. Note: ReLU layers were omitted from description simplicity.

Table 4.7: Model 1

Layer	# Filters	Filter Size
Input	112 × 112	-
Convolutional	32	3
Convolutional	32	3
MaxPolling	2 × 2	2
Convolutional	256	3
Convolutional	256	3
MaxPolling	2 × 2	2
Dense	512	6
Dense	512	1
Output	1	1

Table 4.8: Model 2

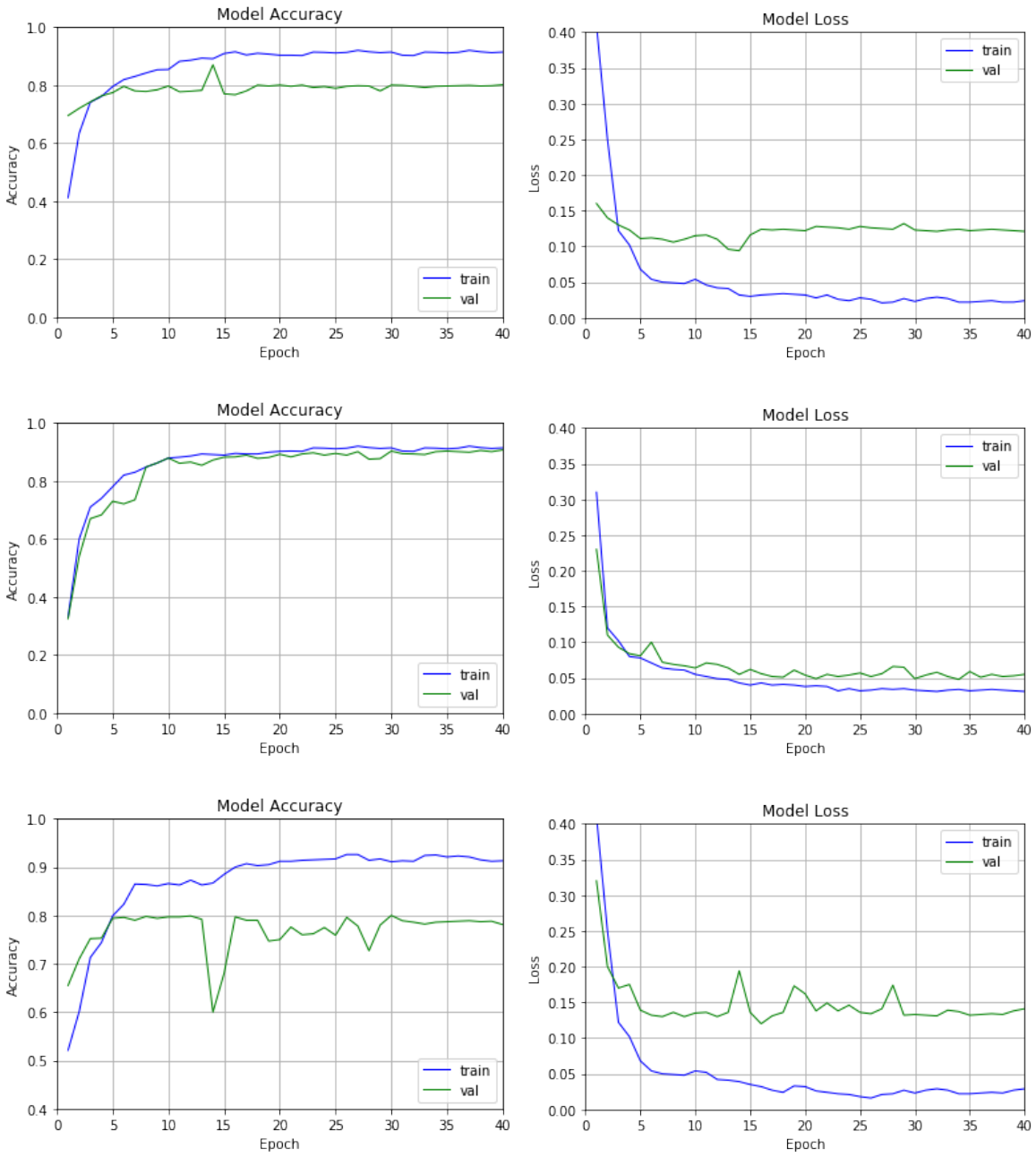
Layer	# Filters	Filter Size
Input	112 × 112	-
Convolutional	96	3
Convolutional	96	3
MaxPolling	2 × 2	2
Convolutional	192	3
Convolutional	192	3
MaxPolling	2 × 2	2
Convolutional	256	3
Convolutional	256	3
Convolutional	256	3
MaxPolling	2 × 2	2
Dense	512	6
Dropout	-	0.5
Dense	512	1
Output	1	1

Table 4.9: Model 3

Layer	# Filters	Filter Size
Input	112 × 112	-
Convolutional	32	3
MaxPolling	2 × 2	2
Convolutional	32	3
MaxPolling	2 × 2	2
Convolutional	64	3
MaxPolling	2 × 2	2
Convolutional	64	3
MaxPolling	64	1
Dense	128	6
Dropout	-	0.5
Dense		1
Output	1	1

Additionally, only training data is subject to augmentation. For this, every patch at training time has an equal probability of being rotated by $90 \times k$ degrees, with $k \in \{0, 1, 2, 3\}$ and horizontal mirroring. *ADAM* was the selected optimizer with constants β_1 and β_2 set to 0.9 and 0.999, respectively. Several experiments were conducted and the final learning rate α was set to 2×10^{-4} and regularization λ to 3×10^{-4}. Models were trained for 40 epochs and batch size was 16. No early stopping was performed. Figures 4.23 presents the loss and

accuracy curves for each of the trained models.



Accuracy Loss

Figure 4.23: Accuracy and loss of the three models individually.

The accuracy obtained for the best model was 0.915 on the test set for Model 2. Globally, high accuracy values were obtained on all models, in part due to the high number of "easy" negatives in this dataset. Although only portions of the breast were considered, some of these regions have low intensity and contrast, making them more obvious to classify. The first model appears to show signs of over-fitting since the difference of accuracy between the

training and validation is increasing in the final epochs. The third model shows a peak in the validation set, suggesting that the learning rate was to higher in this epoch using this architecture and also shows signs of over-fitting.

4.3.5.3 Region Proposal + Classification + Contour refinement

In order to construct a lesion detector, we start by training a CNN network to obtain the heatmap of the lesion regions. Regions containing masses exhibit high values while other regions yield lower values. We can segment an image by making a small modification to an existing classification model to obtain per-pixel class probability.

Considering only values above the defined threshold T, square image patches are extracted from those regions to be classified by a CNN model to reduce the number of False Positives (FP) detections. These two stages enable to identify potential lesion regions and classify them according. On the final stage, positively identified lesions are then subject to contour refinement. In Figure 4.24 summarizes the proposed architecture.

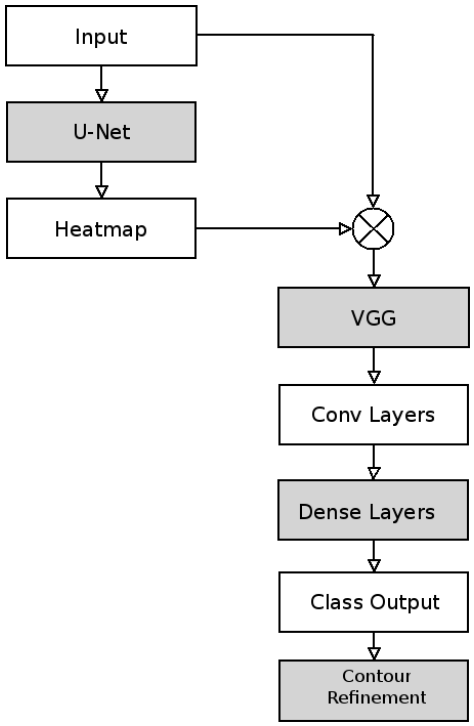


Figure 4.24: Whole Image Screening + Classification Architecture + Contour refinement.

To create the first region proposal stage that corresponds to the mass detection segmentation stage, we make use of a trained CNN and reuse the pre-trained weights and modify last layers, enabling to reuse the trained convolutional layers and adapt the network to our training data as suggested by Xi et al. (2018) and obtain the heatmaps of the lesion.

A deep CNN needs to be cut after the last convolution layer and a global average pooling layer and a fully connected layer must be appended. The new model needs to be re-trained to determine the weights $w_i = (i = 1, 2, \ldots, n)$ for the output layer. The feature maps from the output of the last convolutional layer are denoted as $f_i(i = 1, 2, \ldots, n)$. The importance of the image regions are then identified by projecting back the weights of the output layer onto the convolutional feature maps (Zhou et al., 2016) thought

$$CAM = \sum_{i}^{n} w_i f_i. \tag{4.34}$$

In order to obtain the heatmaps, we make use of the Residual Network (ResNet) (He et al., 2016) and modify the last layers. Since the output layer of the ResNet50 original network is constructed to handle 1000 classes, the output layer of the ResNet50 must be modified to produce a two-class output (mass/background) instead of the 1000 classes by default.

Transfer learning enables to take advantage of the trained feature extraction, by reusing the CNN network architecture and trained weights and fine-tune the model to specific training data without extensive training, and take advantage of the pre-trained feature extractor formed by the convolutional layers. This adaptation commonly consists in removing the last three layers from the network and replaced by three new layers (fully connected layer, soft-max layer, and classification layer) and retrain the new network layers using the training set.

Since ResNet50 has already the required configuration namely the global average polling, we reconfigure the output layer for two class output and retrain the pre-trained network with our training set and perform transfer learning (Hoo-Chang et al., 2016) by freezing the all the convolutional layers except the last one and reuse its weights from Imagenet (Deng et al., 2009) and re-train the network. Global Average Pooling instead contrary to Max Polling that replaces areas with the maximum value, it replaces with the average.

To convert the patch based classifier into a whole image region proposal, we compute the Class Activation Mapping (CAM) for identifying regions of interest on an image using a CNN for the specific class (Zhou et al., 2016), (Figure 4.25). CAM enables to identify image regions relevant to a particular class and at the same time allowing the reuse of pre-trained classifiers for localization purpose. The main advantage of the use of CAM with ResNet50 architecture relies upon the fact that ResNet has already the required architecture and for computing CAM, namely a global average pooling layer, enabling ResNet50 to compute CAM's without further training after the reconfiguration. For class activation map generation we reuse the work of [9].

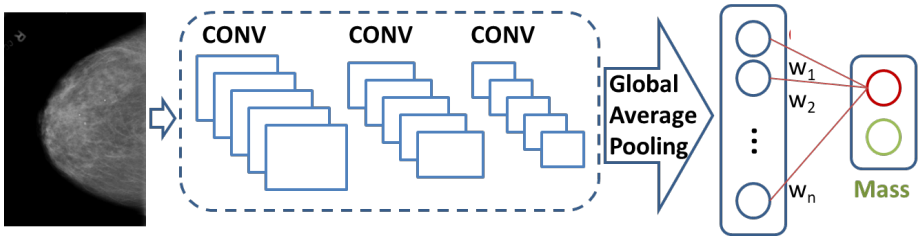


Figure 4.25: Class activation mapping for heatmap production. (Image from Xi et al. (2018))

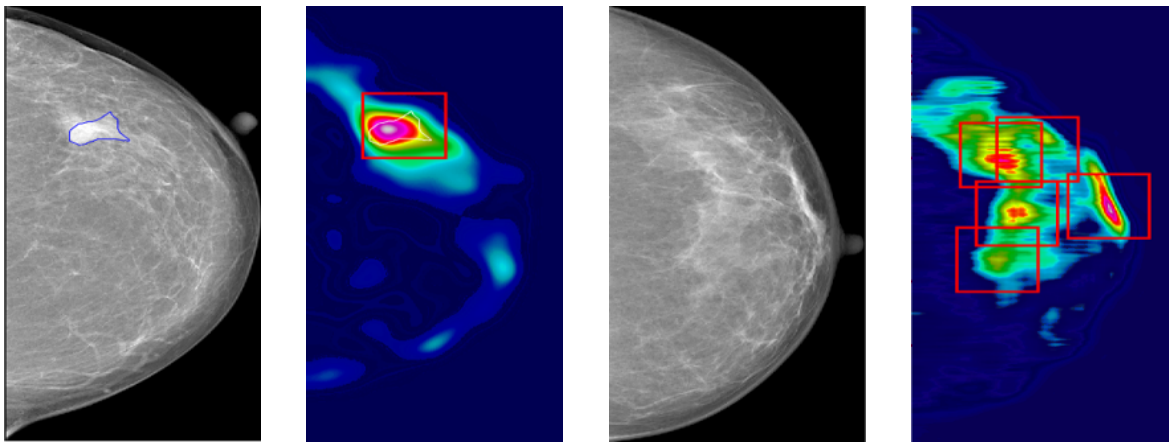
Since patch images in our training set have the size of 112×112 and the network input size is (224×224) with 3 channels, we zero pad the patch images to fit the network original input and replicate the grayscale channel among the three channels of the ResNet50 network. To obtain the probability maps with the same size of the input, the model is exhibited over the whole image and final CAMs are yielded. To avoid memory limitations when using large images a region of 512×512) patches are run each time and outputs concatenated to obtain a final result.

For the network training the following considerations were taken:

- **Augmentation** each patch image and corresponding masks are mirrored and rotated by $0°, 90°, 180°$ and $270°$, increasing the training set by a factor of 8. We crop only the breast region to reduce image area to be screened.
- **Dataset** initially was divided in 80% for training, 20% for testing to determine best parameters, allowing to run a 5-fold cross-validation to determine best parameters during 5 epochs. After determined we set the split into 75% and 25% for train and testing respectively.
- **Loss function** corresponds to the binary cross-entropy for two classes problem.
- **Optimization** was ADAM with ($\beta_1 = 0.9, \beta_2 = 0.995$ and $\epsilon = 10^{-6}$) for optimization. The model was trained for 30 epochs. Batch size was set to 16.
- **Evaluation** we count a breast mass as a true positive if a blob in the segmentation covers at least 20% of its area. The number of FP per image is also computed on the images without a breast mass. $sens_{1FP} = \frac{TP_r}{FP_r}$.
- **Hyper parameter** the final selection using 5-fold cross-validation to determine (α, λ while using the sensitivity at one false positive per image as performance metric. For each fold independently 5 network parameter combinations were trained during 5 epochs and the best final parameters were, $\alpha = 2 \times 10^{-5}$. Then the network was trained during 30 epochs.

[9] https://jacobgil.github.io/deeplearning/class-activation-maps

The new ResNet model trained in the patch images achieved an accuracy of 0.9 on the test set. This value is higher enough to create a reliable screening stage, (Figure 4.26). Images are not subject to any post-processing since we are interested in screening results on this stage.



Example 1 Heat 1 Example 2 Heat 2

Figure 4.26: Pairwise comparison between mammogram image and heapmaps (White - GT).

Regions that exhibit high values in the heatmap corresponds to regions that present high resemblance with masses. However, the pectoral muscle region presented also high values due to intensity and texture similarity. A pre-processing stage like Shortest Path (SP) in polar coordinates can be used to remove pectoral muscle region from images, reducing the number of FP. The second stage corresponds to the classifier itself. Considering only regions that present a mean heatmap value above T, a minimum (112×112) bounding box centered on the weighted heatmap center peak is extracted and subject to the mass patch classifier. All images in the INbreast database test set (410) were subject to region finding, classification and contour refinement by the cascade model.

For base comparison purposes, a state of the art mass lesions detection method proposed by Dhungel et al. (2017) composed by Conditional Random Field (CRF) model with active contour refinement is also listed, yielding a segmentation accuracy of $DICE = 0.850$ while attaining 0.900 of the True Positives (TP). Results comparison are presented in Table 4.10. The mass is considered to be detected if the Intersection over Union (IoU) between the bounding box of the candidate region and ground truth is greater than or equal to 0.2, (Sampat et al., 2005b; te Brake et al., 2000).

Table 4.10: Performance comparison of mass screening detector. Results mean(std).

Author	Setup	Description	Database	DICE
Dhungel et al. (2017)	Min user iter.	CRF model with active contour refinement	INbreast	0.850(0.020)
Dhungel et al. (2015)	Min user iter.	CRF model w/o contour refinement	INbreast	0.900(0.020)
Cardoso et al. (2015)	Manual	SP in Cartesian Coordinates	INbreast	0.880(-)
te Brake et al. (2000)	Manual	Probabilistic method+ radial gradient	Private	0.820(-)

DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better)

Table 4.11 summarizes the results of the first stage implementation compared with the state of the art work (Dhungel et al., 2017) for mass lesion detection. The segmentation metrics are generated by setting each pixel whose CAM value was lower to the threshold T to the background (0) and those whose CAM value is greater than a threshold T to breast mass (255).

Table 4.11: Performance evaluation of mass screening detector. Results mean(std).

Method	FP	TP_r	AOM	CM	DICE
SotA	(-)	0.900(-)	-	-	0.850(0.020)
CNN Screen ($T = 0.6$)	10(1.847)	0.862(0.094)	0.671(0.062)	0.635(0.073)	0.722(0.083)
CNN Screen ($T = 0.8$)	8(1.693)	0.829(0.103)	0.524(0.099)	0.557(0.076)	0.682(0.068)

FP (Number False Positives - lower the better), TP_r=Sens= $\frac{\#TP}{\#TP+\#FN}$ (Detection Rate/Sensibility - higher the better). AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better)

Evaluating the classifier on true positive and false negative responses, when setting a threshold of $T = 0.8$ it attained 0.829 of the TP. This value is slightly lower due to the fact that some extracted images patches were classified as background due to mass center shift and not fully contained mass contour region and some TP were discarded by the selected threshold. When setting the screening threshold to a lower value $T = 0.6$ it attained 0.862 of the TP. This increase is explained by the fact that mass that presents lower heatmap probability value was not discarded and the center was correctly attained by the screening stage, contained its contour and surrounding areas, however with a higher number of FP that were correctly identified as background by the CNN classifier. Table 4.12 presents a summary of the results when of the segmentation stage is fitted with the classifier stage.

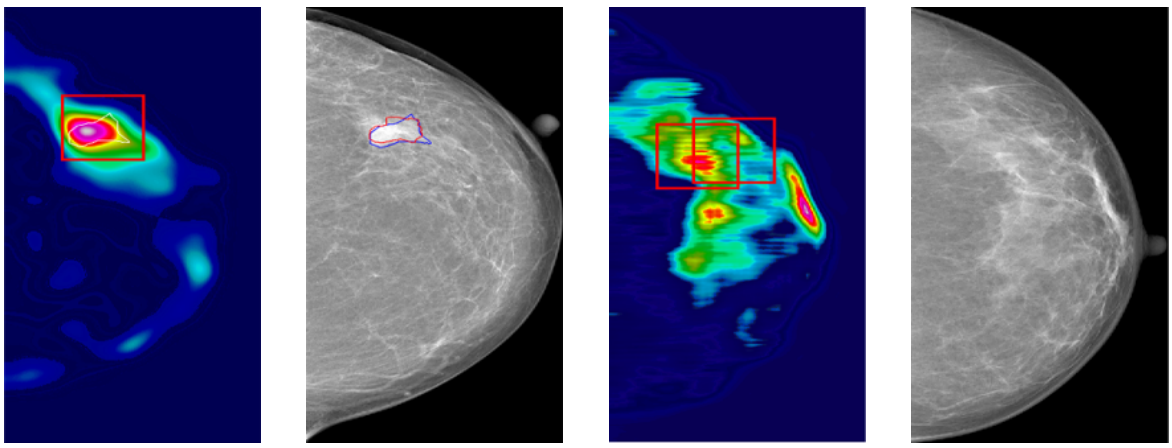
Table 4.12: Performance evaluation of mass screening detector with classifier stage. Results mean(std).

Method	FP	TP_r
SotA.	(-)	0.900(-)
CNN Screen ($T = 0.6$)	10(1.847)	0.862(0.094)
CNN Screen + Class ($T = 0.6$)	3(0.201)	0.853(0.071)
CNN Screen ($T = 0.8$)	8(1.693)	0.829(0.103)
CNN Screen + Class ($T = 0.8$)	2(0.109)	0.762(0.086)

FP (Number False Positives - lower the better), TP_r=Sens= $\frac{\#TP}{\#TP+\#FN}$ (Detection Rate/Sensibility - higher the better), Measures range $[0, 1]$.

The choice of a cascade configuration for breast lesion screening enables to divide the problem into two main components and refine each of the components individually while using small size databases.

To evaluate the performance of the final stage, (contour refinement), the positive identified mass patches are subject to extract contour using the SP in Cartesian Coordinates, described and evaluated on Chapter 3. Figure 4.27 exhibits the final contour extraction after FP have been removed. Results are summarized in Table 4.13 for the two evaluated thresholds, $T = 0.6$ and $T = 0.8$.



Heat 1 Example 1 Heat 2 Example 2

Figure 4.27: Pairwise comparison between mammogram image detections after FP reduction and corresponding contour refinement (GT - Blue, Contour - Red).

Table 4.13: Performance evaluation of mass screening detector + classifier + contour refinement. Results mean(std).

Method	FP	TP_r	AOM	CM	DICE
SotA	(-)	0.900(-)	-	-	0.850(0.020)
CNN + Contour ($T = 0.6$)	3(0.201))	0.853(0.071)	0.719(0.059)	0.712(0.082)	0.829(0.097)
CNN + Contour ($T = 0.8$)	2(0.109)	0.762(0.086)	0.684(0.142)	0.691(0.138)	0.702(0.135)

FP (Number False Positives - lower the better), TP_r=Sens= $\frac{\#TP}{\#TP+\#FN}$ (Detection Rate/Sensibility - higher the better). AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better)

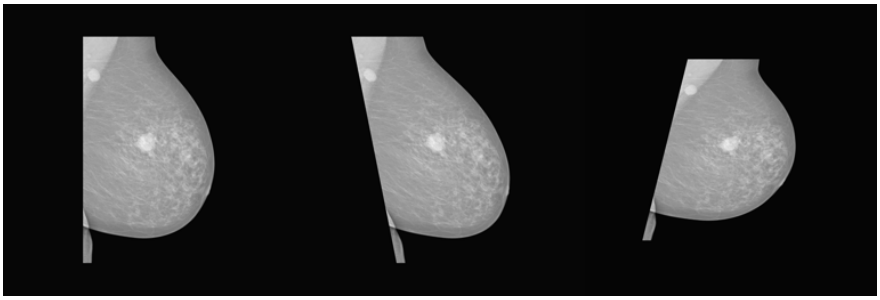
When setting $T = 0.6$ and using the SP in Cartesian Coordinates, image patches were correctly segmented ($DICE = 0.829$) , with results being similar to the manual approach ($DICE = 0.841$). This similarity can be explained by the same image domain of the images. With final contour determined, accessing the malignancy of the findings can be now put in place over the final mass lesion findings to determine the BI-RADS class.

4.3.6 Experiments and Results for Deep Leaning Methods For BI-RADS Classification

In order to construct an image lesion classifier, we start by creating a strong augmented training set in order to increase the robustness of the model. To train the model, we fine tune a pre-trained model to our dataset in order to predict the BI-RADS class.

4.3.6.1 Dataset Construction

For the dataset construction the INbreast database was considered. Images have the maximum the original size of (4084×3328) and to generate the new dataset, the foreground containing the breast region with the pectoral muscle is cropped into a new image and zero padded or scaled if needed into a fixed size of 2048×2048 or scaled to fit if to large. We established the rule of cropping images to have a minimum 10% lateral box relief to perform shear and rotations without cutting boundaries of the breast region. Images were subject to 0.8 up to 1.2 scaling (if possible) with zoom step increases of 0.1 on the cropped area and rotated between $[0°, 90°]$ with an angle interval of $15°$.(Figure 4.28). Random translations and affine transformation of -0.2 to 0.2 in x axis are also performed on each image.



Example Shear Shear H

Figure 4.28: Example of the constructed dataset (Without mirroring).

For training considerations, the constructed dataset is divided into two different subsets with 75% of cases per class being randomly selected for training and remaining for test. All images in the training set are subject to augmentation with each being rotated by $90 \times k$ degrees, with $k \in \{0, 1, 2, 3\}$ and horizontal mirroring. Similar to the traditional machine learning developments, BI-RADS 5 and 6 were merged into a single category. The final number of images are summarized in Table 4.14.

Table 4.14: Databases size per BI-RADS.

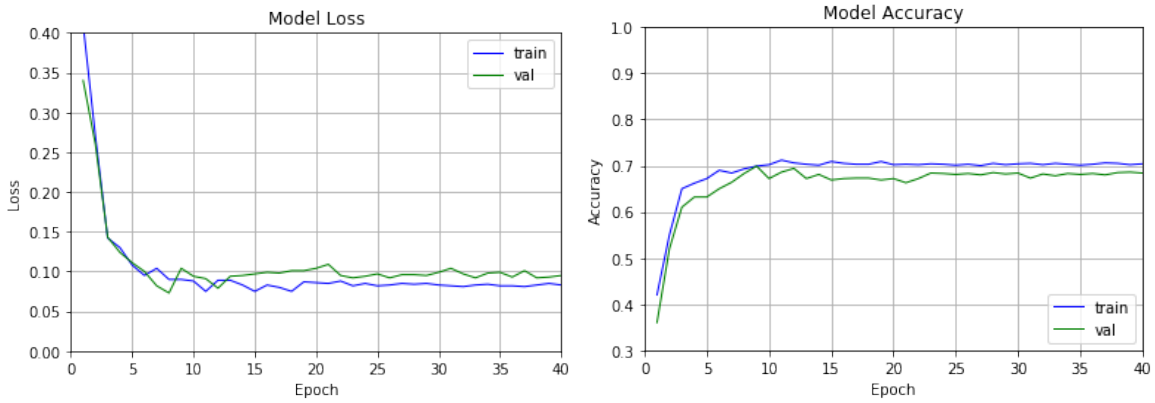
Data	1	2	3	4	56	Total
Orig	67	220	23	43	57	410
Train 75%	50	165	17	32	43	307
Test 25%	17	55	6	10	15	101
Train Aug (A)	250	825	85	160	215	1535
Train Aug (Af + T)	750	2475	255	480	645	4605
Train Aug (A + T + M)	2000	6600	680	1280	1720	12280
Train Aug (A +Af +T + M)	19800	2040	3840	4440	2200	32320

A - Angle, Aff - Affine, T - Translation, M - Mirror

4.3.6.2 Transfer Learning and Training

In order to asses the performance of the CNN networks for ordinal classification, we make use of a pre-trained VGG16 (Simonyan and Zisserman, 2014a), with the original model, also trained in more than a million ImageNet (Deng et al., 2009) as described in Section 4.3.5.3. The choice of VGG16, instead of the ResNet used in the screening stage is due to simplicity. For the task, we make use of transfer learning by freezing the convolution layers and retrain the last convolutional and fully connected layers to classify images into 5 BI-RADS categories. In addition since VGG16 input is a 3 channel images with size 224×224, we also replicate the process carried for the screening and resize our images and replicate the gray image over 3 channels to fit the input of the pre-trained network.

Model was fine tuned during 40 epochs using ADAM with $\alpha = 0.0003$, $\beta_1 = 0.9$ and $\beta_2 = 0.998$ with the loss function corresponding to the categorical cross-entropy with final layer configured to classify 5 different classes. Accuracy was the selected metric, defined as $MaxAcc = \frac{1}{N}\sum_{i=1}^{N} Pred_i \geq True_i$ with N corresponding to the number of examples, and $Pred_i$ and $True_i$ to the predicted and true class respectively.. Training performance are exhibited on Figure 4.29.



Loss - Train/Validation

Acc - Train/Validation

Figure 4.29: Loss and Accuracy during training.

Table 4.15 summarizes the results and MAE class difference are presented in Figure 4.30.

Table 4.15: Attained accuracy in the test set.

Data	MAE
Using Train Aug (A)	1.343(0.503)
Using Train Aug (A+Af+R+M)	0.591(0.013)
Using Train Aug (A+Af+R+M+Z)	0.584(0.011)

MAE (Mean Absolute Error - lower the better), Measures ranging from $[0, 1]$.

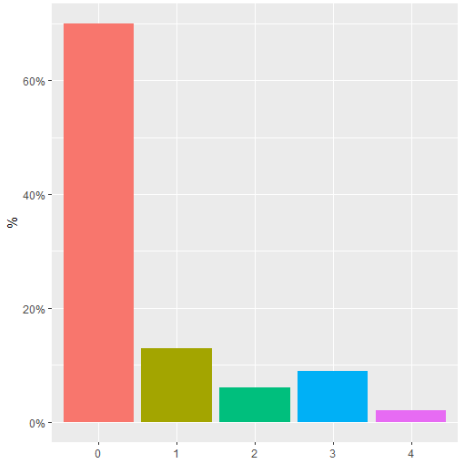


Figure 4.30: MAE class difference distribution.

Using affine transformations for data augmentation combined with rotations enable to increase the accuracy of the model on the test set, proving that data is the main drive motor of the deep learning models. Also, resizing the images to fit our pre-trained network reduces the ability for the model to extract meaningful features from images that contain small lesions manifestations. The MAE was lower since the images contained calcifications and masses or other anomalies were considered as single BI-RADS report, instead of the single mass calcification where the report corresponds to the higher level that may not correspond the mass lesion only.

4.4 Summary

Merging BI-RADS 5 and 6 into single class proved to be a good option to reduce class dispersion, while re-sampling data and addressing ordinality responses improved classification accuracy.

Deep learning approaches to identify and classify lesions proved to be robust since the majority of the lesions were detected and FP where correctly discarded by the classifier stage. In addition, data augmentation proved to be vital to obtain robust models and increase overall accuracy. More classes such as calcification's or pectoral muscle tissue can be included and trained just needing to add appropriate samples images and change the number of output classes by the softmax layer.

Considering deep learning for breast classification the choice of strong augmentation combined with affine transformation enable to increase the accuracy of the model.

Chapter 5

Integrated System Performance

The previous chapters presented a detailed description of each of the components of a Computer Aided Systems (CAD) system complemented with several method pair comparison. In this chapter, the best performer methods are selected for evaluating the performance of the complete system enabling to assess which are the components that have a strong impact on the performance of the system.

5.1 Conducted Experiments

To access the performance of the CAD system in an interactive manner, we conducted the experiment from the beginning of the system towards the end, by replacing the Ground Truth (GT) from the previous stage by its corresponding outputs. Side-by-side evaluation is performed on each block of the fully connected pipeline with its counterpart using the GT information.

5.1.1 Pectoral Muscle Segmentation

Since the pectoral muscle segmentation does not depend on any preceding task, its performance is equal to presented in Section 2.2 (Table 3.4) when selecting the Shortest Path (SP) in polar coordinates.

5.1.2 Mass Lesion Detection and Contour Extraction

For lesion detection and contour extraction we assess the impact of supplying the top performer lesion detection method with outputs of the pectoral muscle segmentation, by comparing the results against those attained by using the GT as that module's input. In the experiment, only images containing masses with or without pectoral muscle were considered. Table 5.1 summarizes the results using the Saliency Map method fitted with a False Positives (FP) reduction stage.

Table 5.1: Performance evaluation of mass lesion detection's with FP rejection with SVM classifier. Results mean (std).

Method	FP	TP_r	AOM	CM	DICE
Saliency(GT)	2(0.094)	0.645(0.084)	0.372(0.069)	0.549(0.089)	0.524(0.097)
Saliency	2(0.095)	0.635(0.086)	0.358(0.071)	0.528(0.091)	0.511(0.107)

FP(False Positives - lower the better), TP_r=Sens= $\frac{\#TP}{\#TP+\#FN}$ (Detection Rate/Sensibility - higher the better). AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better).

Detail analysis showed a small increase in the number of FP. This can be explained by the fact that the perfect muscle region was not totally segmented, leaving areas to be detected by the Saliency Maps that were not discarded by the FP reduction stage. Region metrics remain almost the same.

Considering only detection's that rely inside a mass region and the corresponding GT, image patches were extracted with a 50% increase of the bounding box in both cases for contour extraction. Table 5.2 summarizes the comparisons between mass GT and detection stage when using SP in Cartesian coordinates.

Table 5.2: Performance evaluation of mass lesion contour extraction. Results are in mean (std).

Method	AD	AMED	HD	AOM	CM	DICE
SPCC (GT)	6.824(7.719)	7.655(8.289)	22.148(19.354)	0.729(0.138)	0.836(0.103)	0.841(0.107)
SPCC	14.965(9.321)	19.386(12.429)	32.245(22.564)	0.623(0.265)	0.682(0.203)	0.695(0.275)

AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better), while AD, AMED and HD are measures of pixel error (the lower the better).

As expected the use of the automatic detection's for external contour extraction degraded the system performance. This is due to the under-detected regions combined with center position shifts, difficulting the external final contour determination.

5.1.3 Mass Lesion Classification

To access the impact of the preceding stage in feature extraction and corresponding classification, the extracted and the GT contours are used to evaluate the Breast Imaging Reporting And Data System (BI-RADS) classifier. The same genre of features was calculated and evaluated using the same trained models. The complete pipeline achieved a final Mean Absolute Error (MAE) 0.876 against 0.145 when using the ordinal classifier with re-sampled data. This enables to conclude that the detection stage affects significantly the complete pipeline and efforts must be accomplished to increase its performance.

5.1.4 Overall Results

Overall results are summarized in Table 5.3. It can be concluded that detection is the part of the pipeline that has a higher negative impact on the overall performance of the subsequent blocks.

- Pectoral Muscle Segmentation: No substantial differences where attained since corresponds to the first stage of the pipeline.
- Detection: For the mass detection results, sensitivity and the number of FPs increased when using the automatic pipeline. Following stages were affected in great percentage by under-detected regions.
- Contour Extraction: Suffers significantly from early detection's, namely the under-detection regions that affect significantly contour extraction and consequent feature extraction.
- Feature Extraction and Classification: The BI-RADS assessment suffer from all accumulated error from previous tasks.

Table 5.3: Comparative Analysis of each block.

Stage	GT	Automatic
Muscle Segmentation (SPPC)	AD =0.062(0.021 AMED=0.065(0.015) HD =0.161(0.029) AOM=0.735(0.036) CM =0.822(0.021) DICE=0.799(0.028)	
Detection (Sal + FP Red)	FP =2(0.094) TPr =0.645(0.084) AOM =0.372(0.069) CM = 0.549(0.089) DICE= 0.524(0.097)	FP = 2(0.095) TPr = 0.635(0.086) AOM = 0.358(0.071) CM = 0.528(0.091) DICE= 0.511(0.107)
Contour (SPCC)	AD = 6.824(7.719) AMED= 7.655(8.289) HD =22.148(19.354) AOM = 0.792(0.138) CM = 0.836(0.103) DICE= 0.841(0.107)	AD =14.965(9.321) AMED=19.386(12.429) HD =32.245(22.564) AOM = 0.623(0.265) CM = 0.682(0.203) DICE= 0.695(0.275)
BI-RADS Class (Ordinal)	MAE=0.842	MAE=1.251

AOM, CM and DICE are measures of accuracy ranging from $[0, 1]$ (the higher the better), while AD, AMED and HD are measures of pixel error (the lower the better). FP(False Positives - lower the better), TP_r=Sens= $\frac{\#TP}{\#TP+\#FN}$ (Detection Rate/Sensibility - higher the better).

We can observe that the detection stage deteriorates by a large percentage the accuracy of the final classifier. This is due to under-detected regions that create lesion contour artifacts.

Chapter 6

Conclusions

The current work describes an effort to study and analysis of a framework to guide radiologist in the analysis of mammograms images. A summary of the studied and developed techniques is presented in this chapter and directions for future work are suggested.

6.1 Summary Of Results

In terms of the work, according to the past chapters, it can be divided into different phases, including: (1) pre-processing, (2) detection of suspicious regions and characterization (3) feature extraction and classification. All the experiments were conducted using the INbreast and BCDR database formed by full-field digital mammogram images that, along with the images, contains meta-data information like breast density, Breast Imaging Reporting And Data System (BI-RADS) assessment and Region of Interest (ROI) considered as suspicious regions. Adequate image manipulation can have a strong impact on the performance of the subsequent task. As described, typical pre-processing applied to mammogram images focus in the removal of unwanted regions, namely the removal of artifacts and pectoral muscle. (1) In pre-processing, namely in the pectoral muscle segmentation, five methods were compared, namely Regions Growing, Active contours, intensity based, Shortest Path (SP) in polar coordinates and semantic segmenting using deep learning approaches. All methods were evaluated using thee regions metrics, namely the Area Overlap Measure (AOM), a Combined Measure (CM) of under-segmentation, over-segmentation and Dice Coefficient (DC). For contour error metrics Average Distance (AD), Average Median Distance (AMED) and Hausdorff Distance (HD) was used. Concerning the best performer method, SP in

polar coordinates yielded $AD = 0.062$, $AMED = 0.065$, $HD = 0.161$, $AOM = 0.735$, $CM = 0.822$, $DICE = 0.799$.

Concerning (2), detection of suspicious regions and characterization, the two most common findings in mammogram images are masses and calcification's. For mass detection, three approaches were compared, watershead, saliency maps and Iris filter followed by a False Positives (FP) reduction stage to remove false detection's. The final detection's are subject to a final closed contour segmentation to extract the external contour and obtain the ROI of the lesion. Five contour extraction methods were compared and regions metrics are extracted and evaluated regarding Ground Truth (GT). In the detection's, Saliency map presented the lower FP rate $FP = 5$ and a True positive rate $TP_r = 0.673$, attaining the following region metrics, $AOM = 0.396$, $CM = 0.560$, $DICE = 0.520$.

Considering the FP reduction stage the The FP reduction stage enabled to reduce the FP rate in almost 30%, while the regions metrics were almost maintained. The results are $FP = 2$ $TP_r = 0.645$, $AOM = 0.372$, $CM = 0.549$, $DICE = 0.524$. The TP_r lowered in small portion due to True Positives (TP) being removed by the FP reduction stage.

For lesion contour extraction, SP in Polar and the Cartesian coordinates, Snakes, Gradient Convergence Filters with regularization (SBF-Reg) were considered. SP in the Cartesian coordinates, yielded the best results, with $AD = 6.824$, $AMED = 7.655$, $HD = 22.148$, $AOM = 0.792$, $CM = 0.836$, $DICE = 0.841$. This result can be partially explained by the use of the image in the original coordinates avoiding deformations introduced by the polar transformation. Calcifications were also addressed in this work, namely by the use of a simple outlier detection and Top Hat + wavelet decomposition to extract suspicious ROIs regions that can be related these manifestations.

Feature extraction and classification (3), a review of features used in the literature was presented. For feature selection, a large number of existing features were studied using Principal Component Analysis (PCA) and order of importance. For masses characterization's, only 18 features survived the correlation criteria. For the classification task, two main methodologies were employed, binary and BI-RADS classifications. The use of BI-RADS classification enabled to reveal in detail which are the classes that most contribute to the decrease of accuracy and confusion several combinations of models were evaluated. The ensemble with an LR stacker was the best performer, yielding an accuracy of 0.835. Considering the BI-RADS classification task, the class imbalance proved to have a strong impact on the model capabilities. The conducted experiments enabled to conclude that data augmentation, class, and ordinality influences significantly the model's performance. Merging BI-RADS 5 and 6 classes into a single one managed to reduce the Mean Absolute Error (MAE) from 1.912 to 1.071 using Extreme Randomized Trees (ERT) due to being fitted with bootstrap re-sampling. The re-sampling with Random Forest (RF) enable to decrease further the MAE to 0.891. Considering ordinal classifier using Ordinal Regression

Trees (ORT), the MAE decreased to 0.842 using the re-sampled data, proving that data re-sampling and ordinality play a major role in models performance, and when addressed enable to reduce in a great amount the inter-class miss-classification.

Considering the deep learning approaches, a vast dataset was constructed from patches of the breast regions to train a Convolutional Neural Networks (CNN) model to perform inference in similar patches. The patch models yielded good results (accuracy ≈ 0.900). Combining the the best performer classifier model with a screen stage, a cascade lesion screening and classification was constructed. The methodology consisted in the introduction of dilated convolution to identify regions of interest. Regions detection's above a certain threshold were considered to be classified by the patch previous classifier. The result showed that the cascade enable to reduce the mean number of FP rate by large amount, from 10 to 3 when using a threshold $T = 0.6$, while attaining 0.853 of the TP. The two stages can benefit by being integrated into a single one to automatically determine the best probability threshold. Final contour refinement presented similar results when compared with stand alone using the patch image of the mass. This is due to the same domain of the images, since they are FFDM.

6.2 Future Work

Each of the chapters can be object of further study and development. In what concerns the pectoral muscle segmentation, semantic segmentation with the used of deep learning technique's can be further studied since its simple and can be generalized to other networks easily. Also, making combinations with a more wider database that combines all the public available repositories to give insights about the generalization capability of the pectoral muscle segmentation.

In lesions detection's, calcification's presented a higher sensitivity when compared to masses, however with a higher cost regarding the number of FPs. Architectural distortion's are, according to the literature, the third most common manifestations of breast cancer. It consist in a distortion of the parenchymal architecture without a concomitant mass and its one of the most challenging to detect and classify. Further developments on the detection stage must be performed, namely the combinations of deep learning approaches with some proved segmenting methods to achieve a higher accuracy on detect lesions with an lower FP number.

Concerning feature extraction and classification in deep networks, different architectures in deep leaning method can be implemented, namely the use of Regional CNN (Faster/Mask R-CNN) (Girshick, 2015) for object detection and classification. Also increasing the size of the dataset by combining all public available datasets into a single one enables to models

to generalize well on new cases. Also creation of common GT ground references among the databases such as anomaly description and corresponding GT masks facilitate the use and training of larger models and evaluate different architectures. In addition, the final extracted contour can be subject to classification using the previous extracted features and trained models to assess the malignancy of the findings.

References

Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Süsstrunk. Salient region detection and segmentation. In *International conference on computer vision systems*, pages 66–75. Springer, 2008.

Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

Praful Agrawal, Mayank Vatsa, and Richa Singh. Saliency based mass detection from screening mammograms. *Signal Processing*, 99:29–47, 2014.

Farhan Akram, Jeong Heon Kim, Inteck Whoang, and Kwang Nam Choi. A preprocessing algorithm for the cad system of mammograms using the active contour method. *Applied Medical Informatics*, 32(2):1, 2013.

Amir A Amini, Terry E Weymouth, and Ramesh C Jain. Using dynamic programming for solving variational problems in vision. *IEEE Transactions on pattern analysis and machine intelligence*, 12(9):855–867, 1990.

Edward Azavedo, Sophia Zackrisson, Ingegerd Mejàre, and Marianne Heibert Arnlind. Is single reading with computer-aided detection (cad) as good as double reading in mammography screening? a systematic review. *BMC medical imaging*, 12(1):22, 2012.

Serge Beucher and Fernand Meyer. The morphological approach to segmentation: the watershed transformation. *Optical Engineering-New York-Marcel Dekker Incorporated-*, 34:433–433, 1992.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Helen Blumen, Kathryn Fitch, and Vincent Polkus. Comparison of treatment costs for breast cancer, by tumor stage and type of service. *American health & drug benefits*, 9(1):23, 2016.

D Brzakovic, XM Luo, and P Brzakovic. An approach to automated detection of tumors in mammograms. *IEEE Transactions on Medical Imaging*, 9(3):233–241, 1990.

Jaime S Cardoso, Inês Domingues, Igor Amaral, Inês Moreira, Pedro Passarinho, João Santa Comba, Ricardo Correia, and Maria J Cardoso. Pectoral muscle detection in

mammograms based on polar coordinates and the shortest path. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 4781–4784. IEEE, 2010.

Jaime S Cardoso, Inês Domingues, and Hélder P Oliveira. Closed shortest path in the original coordinates with an application to breast cancer. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(01):1555002, 2015.

Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer, 2015.

Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

Imen Cheikhrouhou, Khalifa Djemal, D Sellami, Hichem Maaref, and Nabil Derbel. New mass description in mammographies. In *Image Processing Theory, Tools and Applications, 2008. IPTA 2008. First Workshops on*, pages 1–5. IEEE, 2008.

Charles K Chui. *An introduction to wavelets*. Elsevier, 2016.

Filipe R Cordeiro, Wellington P Santos, and Abel G Silva-Filho. A semi-supervised fuzzy growcut algorithm to segment and classify regions of interest of mammographic images. *Expert Systems with Applications*, 65:116–126, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Kamila Czaplicka, Helene Włodarczyk, et al. Automatic breast-line and pectoral muscle segmentation. *Schedae Informaticae*, 2011(Volume 20):195–209, 2012.

Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(3):131–156, 1997.

Ingrid Daubechies. *Different perspectives on wavelets*, volume 47. American Mathematical Soc., 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

Nilanjan Dey, Vikrant Bhateja, and Aboul Ella Hassanien. *Medical Imaging in Clinical Applications*. Springer, 2016.

Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. Automated mass detection in mammograms using cascaded deep learning and random forests. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, pages 1–8. IEEE, 2015.

Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical image analysis*, 37:114–128, 2017.

Inês Domingues, Jaime S Cardoso, Igor Amaral, Inês Moreira, Pedro Passarinho, João Santa Comba, Ricardo Correia, and Maria J Cardoso. Pectoral muscle detection in mammograms based on the shortest path with endpoints learnt by svms. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 3158–3161. IEEE, 2010.

Alfonso Rojas Domínguez and Asoke K Nandi. Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. *Pattern Recognition*, 42(6):1138–1148, 2009.

Carl J D'Orsi. *ACR BI-RADS atlas: breast imaging reporting and data system*. American College of Radiology, 2013.

Richard Drake, A Wayne Vogl, and Adam WM Mitchell. *Gray's Anatomy for Students E-Book*. Elsevier Health Sciences, 2009.

Stuart E Dreyfus. An appraisal of some shortest-path algorithms. *Operations research*, 17 (3):395–412, 1969.

Anastasia Dubrovina, Pavel Kisilev, Boris Ginsburg, Sharbell Hashoul, and Ron Kimmel. Computational mammography using deep neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):243–247, 2018.

Matthias Elter, Christian Held, and Thomas Wittenberg. Contour tracing for segmentation of mammographic masses. *Physics in Medicine & Biology*, 55(18):5299, 2010.

Daniel Rodrigues Ericeira, AristóFanes CorrêA Silva, Anselmo Cardoso De Paiva, and Marcelo Gattass. Detection of masses based on asymmetric regions of digital bilateral mammograms using spatial description with variogram and cross-variogram functions. *Computers in biology and medicine*, 43(8):987–999, 2013.

Tiago Esteves, Pedro Quelhas, Ana Maria Mendonça, and Aurélio Campilho. Gradient convergence filters and a phase congruency approach for in vivo cell nuclei detection. *Machine Vision and Applications*, 23(4):623–638, 2012.

Ricardo J Ferrari, Rangaraj M Rangayyan, JE Leo Desautels, RA Borges, and Annie France Frere. Automatic identification of the pectoral muscle in mammograms. *IEEE transactions on medical imaging*, 23(2):232–245, 2004.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156. Springer, 2001.

Karthikeyan Ganesan, U Rajendra Acharya, Chua Kuang Chua, Lim Choo Min, K Thomas Abraham, and Kwan-Hoong Ng. Computer-aided breast cancer detection using mammograms: a review. *IEEE Reviews in biomedical engineering*, 6:77–98, 2013a.

Karthikeyan Ganesan, U Rajendra Acharya, Kuang Chua Chua, Lim Choo Min, and K Thomas Abraham. Pectoral muscle segmentation: a review. *Computer methods and programs in biomedicine*, 110(1):48–57, 2013b.

Andrea Gavlasová, Aleš Procházka, and Martina Mudrová. Wavelet based image segmentation. In *Proc. of the 14th Annual Conference Technical Computing, Prague*, 2006.

Maryellen L Giger. Medical imaging and computers in the diagnosis of breast cancer. In *Photonic Innovations and Solutions for Complex Environments and Systems (PISCES) II*, volume 9189, page 918908. International Society for Optics and Photonics, 2014.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Pelin Gorgel, Ahmet Sertbas, and Osman N Ucan. A wavelet-based mammographic image denoising and enhancement with homomorphic filtering. *Journal of medical systems*, 34 (6):993–1002, 2010.

Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4): 532–550, 1987.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Boulehmi Hela, Mahersia Hela, Hamrouni Kamel, Boussetta Sana, and Mnif Najla. Breast cancer detection: A review on mammograms analysis techniques. In *Systems, Signals & Devices (SSD), 2013 10th International Multi-Conference on*, pages 1–6. IEEE, 2013.

Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

Shin Hoo-Chang, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285, 2016.

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification. 2003.

Yi-Jhe Huang, Ding-Yuan Chan, Da-Chuan Cheng, Yung-Jen Ho, Po-Pang Tsai, Wu-Chung Shen, and Rui-Fen Chen. Automated feature set selection and its application to mcc identification in digital mammograms for breast cancer detection. *Sensors*, 13(4):4855–4875, 2013.

Daniel P. Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

Judy Kilday, Francesco Palmieri, and Martin D Fox. Classifying mammographic lesions using computerized image analysis. *IEEE transactions on medical imaging*, 12(4):664–669, 1993.

Yeong-Taeg Kim. Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE transactions on Consumer Electronics*, 43(1):1–8, 1997.

Hidefumi Kobatake and Shigeru Hashimoto. Convergence index filter for vector fields. *IEEE Transactions on Image Processing*, 8(8):1029–1038, 1999.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Miroslav Kubat, Robert C Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215, 1998.

Paul Kube. Properties of energy edge detectors. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 586–591. IEEE, 1992.

Yihua Lan, Haozheng Ren, and Jinxin Wan. A hybrid classifier for mammography cad. In *Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on*, pages 309–312. IEEE, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Joonwhoan Lee, Suresh Raj Pant, and Hee-Sin Lee. An adaptive histogram equalization based local technique for contrast preserving image enhancement. *International Journal of Fuzzy Logic and Intelligent Systems*, 15(1):35–44, 2015.

R Sawyer Lee, F Gimenez, A Hoogi, and D Rubin. Curated breast imaging subset of ddsm. *The Cancer Imaging Archive*, 2016.

Yanfeng Li, Houjin Chen, Yongyi Yang, and Na Yang. Pectoral muscle segmentation in mammograms based on homogenous texture and intensity deviation. *Pattern Recognition*, 46(3):681–691, 2013.

Chen-Chung Liu, Chung-Yen Tsai, Jui Liu, Chun-Yuan Yu, and Shyr-Shen Yu. A pectoral muscle segmentation algorithm for digital mammograms using otsu thresholding and multiple regression analysis. *Computers & Mathematics with Applications*, 64(5):1100–1107, 2012.

Li Liu, Jian Wang, and Tianhui Wang. Breast and pectoral muscle contours detection based on goodness of fit measure. In *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on*, pages 1–4. IEEE, 2011.

Xiaoming Liu and Jinshan Tang. Mass classification in mammograms using selected geometry and texture features, and a new svm-based feature selection method. *IEEE Systems Journal*, 8(3):910–920, 2014.

MA Guevara Lopez, NG Posada, Daniel C Moura, Raúl Ramos Pollán, José M Franco Valiente, César Suárez Ortega, M Solar, G Diaz-Herrero, IMAP Ramos, J Loureiro, et al. Bcdr: a breast cancer digital repository. In *15th International Conference on Experimental Mechanics*, 2012.

Indra Kanta Maitra, Sanjay Nag, and Samir Kumar Bandyopadhyay. Technique for preprocessing of digital mammogram. *Computer methods and programs in biomedicine*, 107(2):175–188, 2012.

Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

A Mencattini, M Salmeri, P Casti, ML Pepe, F Mangieri, and A Ancona. Local active contour models and gabor wavelets for an optimal breast region segmentation. *Int J Comput Assist Radiol Surg*, 7(1):256–257, 2012.

Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

Dmytro Mishkin, Nikolay Sergievskiy, and Jiri Matas. Systematic evaluation of cnn advances on the imagenet. *arXiv preprint arXiv:1606.02228*, 2016.

Mario Molinara, Claudio Marrocco, and Francesco Tortorella. Automatic segmentation of the pectoral muscle in mediolateral oblique mammograms. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pages 506–509. IEEE, 2013.

Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

Eric N Mortensen and William A Barrett. Interactive segmentation with intelligent scissors. *Graphical models and image processing*, 60(5):349–384, 1998.

Daniel C Moura and Miguel A Guevara López. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *International journal of computer assisted radiology and surgery*, 8(4):561–574, 2013.

Naga R Mudigonda, R Rangayyan, and JE Leo Desautels. Gradient and texture analysis for the classification of mammographic masses. *IEEE transactions on medical imaging*, 19 (10):1032–1043, 2000.

R Nithya and B Santhi. Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *International journal of computer applications*, 28 (6):21–25, 2011.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

Carlos S Pereira, Hugo Fernandes, Ana Maria Mendonça, and Aurélio Campilho. Detection of lung nodule candidates in chest radiographs. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 170–177. Springer, 2007.

Danilo Cesar Pereira, Rodrigo Pereira Ramos, and Marcelo Zanchetta Do Nascimento. Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer methods and programs in biomedicine*, 114(1):88–101, 2014.

Noel Pérez Pérez. Improving variable selection and mammography-based machine learning classifiers for breast cancer cadx. 2015.

Kersten Petersen, Mads Nielsen, Pengfei Diao, Nico Karssemeijer, and Martin Lillholm. Breast tissue segmentation and mammographic risk scoring using deep learning. In *International Workshop on Digital Mammography*, pages 88–94. Springer, 2014.

Judith MS Prewitt. Object enhancement and extraction. *Picture processing and Psychopictorics*, 10(1):15–19, 1970.

Foster J Provost, Tom Fawcett, Ron Kohavi, et al. The case against accuracy estimation for comparing induction algorithms. In *ICML*, volume 98, pages 445–453, 1998.

Giulia Rabottino, Arianna Mencattini, Marcello Salmeri, Federica Caselli, and Roberto Lojacono. Mass contour extraction in mammographic images for breast cancer identification. In *16th IMEKO TC4 Symposium, Exploring New Frontiers of Instrumentation and Methods for Electrical and Electronic Measurements, Florence, Italy*, 2008.

Rangaraj M Rangayyan, Naga R Mudigonda, and JE Leo Desautels. Boundary modelling and shape analysis methods for classification of mammographic masses. *Medical and Biological Engineering and Computing*, 38(5):487–496, 2000.

Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38(1):35–44, 2004.

Alfonso Rojas-Domínguez and Asoke K Nandi. Development of tolerant features for characterization of masses in mammograms. *Computers in Biology and Medicine*, 39(8): 678–688, 2009.

Larissa CS Romualdo, Marcelo AC Vieira, Homero Schiabel, Nelson DA Mascarenhas, and Lucas R Borges. Mammographic image denoising and enhancement using the anscombe transformation, adaptive wiener filtering, and the modulation transfer function. *Journal of digital imaging*, 26(2):183–197, 2013.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Paul L Rosin. Classification of pathological shapes using convexity measures. *Pattern Recognition Letters*, 30(5):570–578, 2009.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

Nafiza Saidin, Harsa Amylia Mat Sakim, Umi Kalthum Ngah, and Ibrahim Lutfi Shuaib. Segmentation of breast regions in mammogram based on density: a review. *arXiv preprint arXiv:1209.5494*, 2012.

Ms SM Salve and VA Chakkarwar. Classification of mammographic images using gabor wavelet and discrete wavelet transform. *International Journal of Advanced Research in Electronics and Communication Engineering*, 2(5):pp–573, 2013.

Mehul P Sampat, Alan Conrad Bovik, and Mia K Markey. Classification of mammographic lesions into bi-rads shape categories using the beamlet transform. In *Medical Imaging 2005: Image Processing*, volume 5747, pages 16–26. International Society for Optics and Photonics, 2005a.

Mehul P Sampat, Mia K Markey, Alan C Bovik, et al. Computer-aided detection and diagnosis in mammography. *Handbook of image and video processing*, 2(1):1195–1217, 2005b.

Tiago André Guedes Santos. Weighted multiple kernel learning for breast cancer diagnosis applied to mammograms. 2017.

PK Saranya and ES Samundeeswari. A study on morphological and textural features for classifying breast lesion in ultrasound images. *Int J Innov Res Sci Eng Technol*, 5:3267–3279, 2016.

Li Shen. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *arXiv preprint arXiv:1708.09427*, 2017.

SeungYeon Shin, Soochan Lee, and Il Dong Yun. Classification based micro-calcification detection using discriminative restricted boltzmann machine in digitized mammograms. In *Medical Imaging 2014: Computer-Aided Diagnosis*, volume 9035, page 90351L. International Society for Optics and Photonics, 2014.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014a.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014b.

Enmin Song, Shengzhou Xu, Xiangyang Xu, Jianye Zeng, Yihua Lan, Shenyi Zhang, and Chih-Cheng Hung. Hybrid segmentation of mass in mammograms using template matching and dynamic programming. *Academic radiology*, 17(11):1414–1424, 2010.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Gábor J Székely and Maria L Rizzo. Brownian distance covariance. *The annals of applied statistics*, pages 1236–1265, 2009.

Saeid Asgari Taghanaki, Yonghuai Liu, Brandon Miles, and Ghassan Hamarneh. Geometry-based pectoral muscle segmentation from mlo mammogram views. *IEEE Transactions on Biomedical Engineering*, 64(11):2662–2671, 2017.

Maxine Tan, Jiantao Pu, and Bin Zheng. Optimization of breast mass classification using sequential forward floating selection (sffs) and a support vector machine (svm) model. *International journal of computer assisted radiology and surgery*, 9(6):1005–1020, 2014.

Xiaoou Tang. Texture information in run-length matrices. *IEEE transactions on image processing*, 7(11):1602–1609, 1998.

Yimo Tao, Shih-Chung B Lo, Matthew T Freedman, Erini Makariou, and Jianhua Xuan. Automatic categorization of mammographic masses using bi-rads as a guidance. In *Medical Imaging 2008: Computer-Aided Diagnosis*, volume 6915, page 691526. International Society for Optics and Photonics, 2008.

Guido M te Brake, Nico Karssemeijer, and Jan HCL Hendriks. An automatic method to discriminate malignant masses from normal tissue in digital mammograms1. *Physics in Medicine & Biology*, 45(10):2843, 2000.

MIT Technology. The machines are getting ready to play doctor, 2017. URL `https://www.technologyreview.com/s/608234/the-machines-are-getting-ready-to-play-doctor/`.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Hamid R Tizhoosh, MehrdadJ Gangeh, Hadi Tadayyon, and Gregory J Czarnota. Tumour roi estimation in ultrasound images via radon barcodes in patients with locally advanced breast cancer. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1185–1189. IEEE, 2016.

Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108, 2015.

Giovanni Trovini, Christian Napoli, Robert Marti, Amaya Martin, Alessandro Bria, Claudio Marrocco, Mario Molinara, Francesco Tortorella, and Oliver Diaz. A deep learning framework for micro-calcification detection in 2d mammography and c-view. In *14th International Workshop on Breast Imaging (IWBI 2018)*, volume 10718, page 1071811. International Society for Optics and Photonics, 2018.

A Vadivel and B Surendiran. A fuzzy rule-based approach for characterization of mammogram masses into bi-rads shape categories. *Computers in biology and medicine*, 43(4):259–267, 2013.

Marina Velikova, Maurice Samulski, Peter JF Lucas, and Nico Karssemeijer. Improved mammographic cad performance using multi-view information: a bayesian network framework. *Physics in Medicine & Biology*, 54(5):1131, 2009.

Hongting Wang, Jun-Bao Li, Ligang Wu, and Huijun Gao. Mammography visual enhancement in cad-based breast cancer diagnosis. *Clinical imaging*, 37(2):273–282, 2013.

Jinhua Wang, Xi Yang, Hongmin Cai, Wanchang Tan, Cangzheng Jin, and Li Li. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific reports*, 6:27327, 2016.

Chia-Hung Wei, Yue Li, and Pai Jung Huang. Mammogram retrieval through machine learning within bi-rads standards. *Journal of biomedical informatics*, 44(4):607–614, 2011.

Jun Wei, Yoshihiro Hagihara, and Hidefumi Kobatake. Detection of cancerous tumors on chest x-ray images-candidate detection filter and its evaluation. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 3, pages 397–401. IEEE, 1999.

Fred Winsberg, Milton Elkin, Josiah Macy Jr, Victoria Bordaz, and William Weymouth. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*, 89(2):211–215, 1967.

Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

Pengcheng Xi, Chang Shu, and Rafik Goubran. Abnormality detection in mammography using deep convolutional neural networks. *arXiv preprint arXiv:1803.01906*, 2018.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 815–824. ACM, 2006.

Erhu Zhang, Fan Wang, Yongchao Li, and Xiaonan Bai. Automatic detection of microcalcifications using mathematical morphology and a support vector machine. *Biomedical materials and engineering*, 24(1):53–59, 2014.

Xiaoyong Zhang, Noriyasu Homma, Shotaro Goto, Yosuke Kawasumi, Tadashi Ishibashi, Makoto Abe, Norihiro Sugita, and Makoto Yoshizawa. A hybrid image filtering method for computer-aided detection of microcalcification clusters in mammograms. *Journal of Medical Engineering*, 2013, 2013.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

Y-T Zhou, Rama Chellappa, Aseem Vaid, and B Keith Jenkins. Image restoration using a neural network. *IEEE transactions on acoustics, speech, and signal processing*, 36(7): 1141–1151, 1988.

Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. Academic Press Professional, Inc., 1994.

Appendix A

Background Knowledge

A.1 Segmentation Metrics

As the evaluation techniques apply to more than one chapter, they are summarized in the following sections.

A.1.1 Region Based Segmentation Metrics

Region based measures are measures of accuracy (the higher the better). Considering a segmentation problem where T corresponds to the Ground Truth (GT) and S to the automatically segmented Area Overlap Measure (AOM) or Jaccard Index J_i, quantifies the percentage of of the area that relies on the GT in the follow form

$$AOM = J_i = \frac{S \cap T}{S \cup T} \tag{A.1}$$

A Combined Measure (CM) (Elter et al., 2010) is an alternative metric that balances the under-segmentation (U), over-segmentation (O) and AOM, being defined as

$$CM = \frac{AOM + (1 - U) + (1 - O)}{3} \tag{A.2}$$

where

$$U = \frac{|T\ (S \cap T)|}{T}, \qquad O = \frac{|S\ (S \cap T)|}{|S|} \tag{A.3}$$

Sørensen-Dice similarity or Dice Coefficient (DC) , is a statistic used for comparing the similarity of two samples, defined as

$$Dice(S,T) = \frac{|S \cap T|}{\alpha \cdot |S| + (1-\alpha) \cdot T} \tag{A.4}$$

where α can be comprehended between $[0,1]$

$$\begin{cases} \alpha > 0.5 & \text{Precision if more important} \\ \alpha < 0.5 & \text{Precision if more important} \\ \alpha = \frac{1}{2} & \frac{2\,|S \cap T|}{|S|+|T|}. \end{cases} \tag{A.5}$$

The Jaccard coefficient $d_J(S,T)$ measures dissimilarity of the groups, ranging from $[0,1]$, where zero corresponds to minimum distance among both sets and one to the maximum, being defined as

$$d_J(S,T) = 1 - J_i(S,T) = \frac{|S \cup T| - |S \cap T|}{|S \cup T|}. \tag{A.6}$$

.

A.1.2 Contour Based Segmentation Metrics

Contour based measures are measures of error (the lower the better). Hausdorff Distance (HD) (Song et al., 2010), Huttenlocher et al. (1993), also called PompeiuHausdorff distance, measures how far two subsets of a metric space are from each other. It turns the set of non-empty compact subsets of a metric space into a metric space in its own right. In addition, Average Distance (AD) and Average Minimum Euclidean Distance (AMED) can be derived using same base formulation.

$$AD(A,B) = \frac{1}{2}\left[\frac{1}{m}\sum_{i=1}^{m} d(a_i,B) + \frac{1}{n}\sum_{j=1}^{n} d(b_j,A)\right] \tag{A.7}$$

$$AMED(A,B) = \max\left[\frac{1}{m}\sum_{i=1}^{m} d(a_i,B) + \frac{1}{n}\sum_{j=1}^{n} d(b_j,A)\right] \tag{A.8}$$

$$HD(A,B) = \max\left[\max_{i\in 1,\ldots,m} d(a_i,B), \max_{j\in,\ldots,n} d(b_j,A)\right] \tag{A.9}$$

where $A = a_1, a_2, ..., a_m$ and $B = b_1, b_2, ..., b_n$ are the two contours to be compared and $d(a_i; B) = \min_{j \in 1,...,n} ||a_i - b_j||$ is the distance from a_i to the closest point on contour B.

A.2 Model Evaluation Metrics

According to the task (classification or regression) different metrics can be employed to assess the performance of the models.

A.2.1 Classification Metrics

Considering a two-class problem, the confusion matrix of a classifier (Table A.1) reports **i)** the number of instances correctly classified as True Positives (TP) and True Negatives (TN), and **ii)** the wrongly classified instances as Type I errors or False Positives (FP) and Type II errors or False Negatives (FN).

Table A.1: Confusion matrix for two-class classification problem.

		Predicted	
		Positive	Negative
True	Positive	TP	FN
	Negative	FP	TN

Accuracy (Equation A.10) and its complement error rate are standard classification performance metrics which can be extracted from this matrix.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{A.10}$$

In many cases, these metrics are not the most appropriate when the preference is the least frequent class in an imbalanced domain since the minority class has a comparatively smaller impact on the results. In this case, other more appropriate metrics should be used, such as

$$\text{true positive rate (recall, sensitivity or hit rate): } TPR = \frac{TP}{TP + FN} \tag{A.11}$$

$$\text{true negative rate (specificity): } TNR = SPC = \frac{TN}{FP + TN} \tag{A.12}$$

$$\text{false positive rate (fall out): } FPR = \frac{FP}{FP + TN} = 1 - TNR \tag{A.13}$$

$$\text{positive predictive value (precision): } PPV = \frac{TP}{TP + FP} \tag{A.14}$$

Since there is a trade-off between some of these measures and it is impractical to monitor more than one, alternative measures were proposed. The *F-measure* or F_β*-score* (based on Van Rijsbergens effectiveness measure) corresponds to the harmonic mean of precision and recall, attaching β times as much importance to recall as precision (Equation A.15). The *G-mean* (Kubat et al., 1998) is the geometric mean of specificity and sensitivity (Equation A.16).

$$F_\beta = \frac{(1 + \beta)^2 \cdot precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{A.15}$$

$$G\text{-}mean = \sqrt{sensitivity \ \cdot \ specificity} \tag{A.16}$$

The area under a Receiver Operating Characteristic (ROC) curve (AUC) (Metz, 1978; Provost et al., 1998) is yet another popular way to assess the performance of a classifier. Each point of the curve corresponds to the pair (True Positive Rate (TPR), False Positive Rate (FPR)) obtained by using a different decision or threshold parameter for classifying examples.

$$AUC = \frac{1 + TPR - FPR}{2} = \frac{TPR + TNR}{2} \tag{A.17}$$

Kappa statistic (or value) is a metric that compares a Observed Accuracy with a Expected Accuracy (random chance), (Equation A.18).

$$Kappa = \frac{(\text{Observed accuracy} - \text{Expected Accuracy})}{(1 - \text{Expected accuracy})}. \tag{A.18}$$

A.2.2 Regression Metrics

Standard metrics for regression include Mean Square Error (MSE) and Mean Absolute Error (MAE) as defined in Equations A.19 and A.20, where y_i is a true value and $\hat{y}_i$ its prediction.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{A.19}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{A.20}$$

Relative error metrics are unit less which means that their scores can be compared across different domains and are calculated by comparing the scores of the model under evaluation against the scores of some baseline model. The relative score is expected to be a value between $0, 1]$, with values nearer (or even above) 1 representing performances as bad as the baseline model, which is usually chosen as something too naive. The most common baseline model is the constant model consisting of predicting for all test cases the average target variable value calculated in the training data. Normalized Mean Squared Error (NMSE) (Equation A.21) and Normalized Mean Absolute Error (NMAE) (Equation A.22)

$$NMSE = \frac{\sum_{i=1}^{n}(\hat{y} - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2}. \tag{A.21}$$

$$NMAE = \frac{\sum_{i=1}^{n}|\hat{y} - y_i|^2}{\sum_{i=1}^{n}|\bar{y} - y_i|^2}. \tag{A.22}$$

Mean Average Percentage Error(MAPE) (Equation A.23)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i}{y_i}. \tag{A.23}$$

The correlation between the predictions and the true values ($\rho_{\hat{y},y}$) is given by

$$\rho_{\hat{y},y} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{A.24}$$